Missing Data in Cluster Randomised Trials

Md Anower Hossain

Thesis submitted in accordance with the requirements for the degree of Doctor of Philosophy of the University of London
December 2017

Department of Medical Statistics
Faculty of Epidemiology and Population Health
London School of Hygiene and Tropical Medicine

Funded by the Economic and Social Research Council (ESRC), UK

In memory of Razaul Karim & Ahasan Habib

Declaration

I declare that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been acknowledged in the thesis.

Signed:

Date: **December 2017**
Full name: **Md Anower Hossain**

Abstract

Missing outcomes are a commonly occurring problem in cluster randomised trials, which can lead to biased and inefficient inference if ignored or handled inappropriately. Handling missing data in CRTs is complicated due to the hierarchical structure of the data. Two approaches for analysing such trials are cluster-level analysis and individual-level analysis. An assumption regarding missing outcomes in CRTs that is sometimes plausible is that missingness depends on baseline covariates, but conditioning on these baseline covariates, not on the outcome itself, which is known as a covariate dependent missingness (CDM) mechanism. The aim of my thesis was to investigate the validity of the approaches to the analysis of CRTs for the three most common outcome types: continuous, binary and time-to-event, when outcomes are missing under the assumption of CDM. Missing outcomes were handled using complete records analysis (CRA) and multilevel multiple imputation (MMI).

We investigated analytically, and through simulations, the validity of the different combinations of the analysis model and missing data handling approach for each of the three outcome types. Simulations studies were performed considering scenarios depending on whether the missingness mechanism is the same between the intervention groups and whether the covariate effect is the same between the intervention groups in the outcome

model. Based on our analytical and simulations results, we give recommendations for which methods to use when the CDM assumption is thought to be plausible for missing outcomes. The key findings of this thesis are as follows.

Continuous outcomes

- Cluster-level analyses using CRA are in general biased unless the intervention groups have the same missingness mechanism and the same covariate effects on outcome in the data generating model.
- In the case of individual-level analysis, the linear mixed model (LMM) using CRA adjusted for covariates such that the CDM assumption holds gives unbiased estimates of intervention effect regardless of whether the missingness mechanism is the same or different between the intervention groups, and whether there is an interaction between intervention and baseline covariates in the data generating model for outcome, provided that such interaction is included in the model when required.
- There is no gain in terms of bias or efficiency of the estimates using MMI over CRA as long as both approaches use the same functional form of the same set of baseline covariates.

Binary outcomes

- The adjusted cluster-level estimator for estimating risk ratio (RR) using full data is consistent if the true data generating model is a log link model, the functional form of the baseline covariates is the same between the intervention groups, and the random effects distribution is the same between the intervention groups.

- Cluster-level analyses using CRA for estimating risk difference (RD) are in general biased. For estimating RR, cluster-level analyses using CRA are valid if the true data generating model has log link and the intervention groups have the same missingness mechanism and the same functional form of the covariates in the outcome model.

- In contrast, MMI followed by cluster-level analyses gives valid inferences for estimating RD and RR regardless of whether the missingness mechanism is the same or different between the intervention groups, and whether there is an interaction between intervention and baseline covariate in the outcome model, provided that such interaction is included in the imputation model when required.

- In the case of individual-level analysis, both random effects logistic regression (RELR) and generalised estimating equations (GEE) give valid inferences using both CRA (adjusted for covariates such that the CDM assumption holds) and MMI regardless of whether the missingness mechanism is the same or different between the intervention groups, and whether there is an interaction between intervention and baseline covariate in the outcome model, provided that such interaction is included in both the imputation model and the analysis model when required.

- Like continuous outcomes, in the absence of auxiliary variables, there is no benefit in performing MMI rather than doing CRA in terms of bias or efficiency of the estimates.

Time-to-Event outcomes

- In the case of censored data, the unadjusted cluster-level analysis for estimating rate ratio (RaR) is consistent when the event rates are small and the covariate effects are the same between the intervention groups. In contrast, the adjusted cluster-level analysis for estimating RaR is consistent for any event rates when the the covariate effects are the same between the intervention groups.

- The gamma shared frailty model as an individual-level analysis underestimates the standard errors (SEs) of the estimates when each intervention group has small number of clusters.

- The Williams approach performs better than the Greenwood approach for estimating the SEs of Kaplan-Meier (KM) estimates unless the event rate is low and the value of intraclass correlation coefficient is very small.

Acknowledgements

My most heartfelt thanks go to my supervisors **Dr. Jonathan Bartlett** and **Dr. Karla DiazOrdaz** for their excellent supervision, continuous encouragement, valuable suggestions and assistance. It would have never been possible to finish this thesis without their guidance and unlimited times they offered me.

I am very grateful to my associate-supervisor Professor Dr. Elizabeth Allen for her constructive comments and suggestions about this work. I would like to thank Matteo, Schadrac, Kleio, Gaurav and Simon for the time and knowledge we shared together during my study period at London School of Hygiene and Tropical Medicine (LSHTM). I would also like to thank Sarah Thorne, Jenny Fleming and Lauren Dalton in the office of the school for their very cordial and quick response to me.

I am deeply indebted to Professor Syed Shahadat Hossain, University of Dhaka, and Dr. Ranjit Lall, Warwick Medical School, University of Warwick, for their continuous support and encouragement. I am grateful to the University of Dhaka, Bangladesh, for approving my study leave to pursue doctoral degree. I am also grateful to The Economic

and Social Research Council (ESRC), UK, for funding this research via The Bloomsbury Doctoral Training Centre. I would like to acknowledge The Charles Wallace Bangladesh Trust for awarding me a small study grant.

Finally, never enough thanks to my wife Mariam Ratna, my daughter Anisa Hossain and my parents for their continuous support, inspiration and sacrifice during the last four years.

Acronyms and Abbreviations

ABB Approximate Bayesian Bootstrap

CDM Covariate Dependent Missingness

CEA Cost-Effectiveness Analysis

CI Confidence Interval

CRA Complete Records Analysis

CRTs Cluster Randomised Trials

DF Degrees of Freedom

GEE Generalised Estimating Equations

HALI Health and Literacy Intervention

ICC Intraclass Correlation Coefficient

KM Kaplan-Meier

LMM Linear Mixed Model

LOCF Last Observation Carried Forward

MAR Missing At Random

MCAR Missing Completely At Random

MI Multiple Imputation

MLE Maximum Likelihood Estimate

MMI Multilevel Multiple Imputation

MNAR Missing Not At Random

RD Risk Difference

RR Risk Ratio

RaR Rate Ratio

RELR Random Effects Logistic Regression

SBP Systolic Blood Pressure

SE Standard Error

SFM Shared Frailty Model

VIF Variance Inflation Factor

Contents

List of tables

List of figures

Part I

Introduction

Chapter 1

Cluster Randomised Trials

This chapter gives an overview of cluster randomised trials (CRTs). In Section 1.1, we outline a brief introduction to CRTs including advantages over standard trials and the consequences of intraclass correlation in such trials. Section 1.2 explains the most common types of outcome in CRTs. Finally, in Section 1.3, we review the approaches to the analysis of data from CRTs. In the following chapter, we discuss the issue of missing data in CRTs.

1.1 Introduction

Cluster randomised trials (CRTs) are experiments in which clusters of individuals such as villages, schools, or medical practices, rather than individuals, are randomly allocated to intervention and control groups, while individual-level outcomes of interest are observed within each cluster. The number of clusters may vary between control group and intervention group; and the number of individuals in each cluster, known as cluster size, may also vary from cluster to cluster. CRTs with equal number of clusters in each intervention group in addition to constant cluster size are known as balanced CRTs. Examples of CRTs include: (i) communities in a developing country, such as a village or a district, selected as the randomization unit to measure the effectiveness of improved water supplies on childhood diarrhoea, (ii) schools selected as the unit of randomization to evaluate new educational guidelines directed by the ministry of education, and (iii) hospitals selected as the randomization unit to measure the effect of a training program for doctors on the quality of diagnosis and treatment of a specific type of disease. CRTs have been increasingly accepted in the fields of health promotion and health service research by public health researchers. Reasons for this popularity may include the nature of the intervention that itself may dictate its application at the cluster level, less risk of intervention contamination, and greater administrative convenience [1]. However, it is well known that the power and precision of CRTs are lower relative to trials that individually randomise the same number of participants. In spite of having this limitation in terms of statistical power and precision of the parameter estimates, the advantages associated with CRTs are sometimes perceived by researchers to outweigh the resulting cost in statistical power and precision. The reduction in precision due to using CRTs is a function of the variance inflation factor (VIF), also known as the design effect (DF)

[2], which measures how much the sampling variability differs due to clustering from the sampling variability of individual randomisation. In the case of a balanced CRT, the VIF is given by

$$\text{VIF} = 1 + (m - 1)\rho,$$

where m is the constant cluster size and ρ is the intraclass correlation coefficient (ICC) which measures how much more similar the outcomes of individuals in the same cluster are compared to the outcomes of the other clusters. This can also be interpreted as the usual pair-wise correlation coefficient between any two outcomes of the same cluster. The VIF increases with the cluster size and with the intraclass correlation coefficient. The case $\rho = 0$ implies that there is no linear dependency among the individuals in the same cluster. In this case individuals within the same cluster are not more similar compared to the individuals in the others clusters. On the other hand, $\rho = 1$ corresponds to perfect dependence among the individuals in the same cluster. In this case all outcomes of individuals in the same cluster are identical and so the total information provided by a cluster is no more than that provided by an individual in that cluster. In practice, the resulting value of ICC is usually small and typically ranges from 0.001 to 0.05 in primary care and health research; and it is rare to have ICCs above 0.1 [3]. A small value of ICC can lead to substantial VIF and should not be ignored in the design and analysis of CRTs [4].

1.2 Types of outcome in CRTs

In most health service and epidemiological research, the three most common types of outcome are:

- **Continuous outcome**: A quantitative variable either discrete or continuous is measured on each individuals studied in the trial. The difference in mean between control and intervention groups is usually the parameter of interest, although other possibilities may include difference in median or in quartile. An example is the difference in mean number of sexual partners in each group in a sexually transmitted diseases prevention program. The data obtained from the trial are used to estimate the true means of the control and intervention groups and the difference between them.

- **Binary outcome**: occurs when each individual either does or does not satisfy certain criteria. For example, in a trial of a smoking cessation program, the outcome from each individual is either "yes" or "no" depending on whether he/she stopped smoking by the end of the study. The parameter of interest may be risk difference or risk ratio. The risk difference is defined as the difference between the true proportions of individuals who stopped smoking in the control and intervention groups, whereas the risk ratio is defined as the ratio of the true proportions of the control and intervention groups.

- **Time-to-event outcome**: All individuals in the trial are followed until they experience the event of interest or they are censored. The parameter of interest is usually the rate ratio of the event of interest, although one may have interest on difference between the event rates of the control and intervention groups instead. Examples might include the difference between incidence rates of polio per 1000 persons-years of two groups of a polio vaccine trial.

1.3 Analysis of CRTs

As described earlier, outcomes of individuals in the same cluster in CRTs are usually correlated. Standard methods of analysing data from randomised trials assume that the outcomes are statistically independent but this assumption is violated when clusters of individuals are randomised into control and intervention groups. Therefore, special methods are required to analyse CRTs that take into account the correlation between outcomes in the same cluster. The two main approaches to the analysis of CRTs are:

1. Cluster-level analysis, and
2. Individual-level analysis.

1.3.1 Cluster-level analysis

This approach is conceptually very simple as the clusters are the experimental units in CRTs. It is reasonable to obtain a relevant summary measure of the outcome variable for each of these units and to compare these summary measures between the control and intervention groups. This approach can be explained as a two stage process.

In the first stage, a relevant summary measure of the outcome variable is calculated for each cluster based on all outcomes of individuals in that cluster. This might be the mean, proportion or other cluster level statistic. For example, in a trial of systolic blood pressure (SBP) control, a relevant summary measure might be the mean SBP in

each cluster or the proportion of individuals in each cluster who has SBP less than 120 mmHg. The total number of observations in each group is then equal to the number of clusters in that group.

In the second stage, the two sets of cluster specific summary measures obtained in the first stage are compared using appropriate statistical methods. The most common one is the standard $t-$test for two independent samples since the resulting summary measures are statistically independently, which is a consequence of the clusters being independent of each other. The corresponding non-parametric methods could be *Wilcoxon rank sum test* or *permutation test* [5].

The cluster-level approach with $t-$test is robust in terms of type I error and confidence interval with approximately correct coverage [5], but it may not be efficient in terms of precision and power when cluster sizes vary widely which is very common in practice. The reason behind this is that equal weight is given to each cluster-level summary ignoring the variation in cluster sizes. Furthermore, the equal variance assumption may be violated if the cluster sizes vary substantially [5].

In CRTs, baseline covariates that may be related to the outcome of interest are often collected and incorporated into the analysis. These baseline covariates could be measured at cluster-level or individual-level. The main purposes of adjusting for covariates is to increase the credibility of the trial findings by demonstrating that any observed intervention effect is not attributable to the possible imbalance between control and intervention groups in terms of baseline covariates, and to improve the statistical power [6]. Randomisation ensures that the control and intervention groups are balanced on average in terms of baseline covariates. In practice there will be some imbalance by chance, when

the number of clusters is small [5]. This imbalance could, by chance, be quite large. To adjust for baseline covariates in cluster-level analysis, an individual-level regression analysis of the outcome of interest is carried out at the first stage of analysis ignoring the clustering of the data, which incorporates all covariates into the regression model except the intervention indicator [5, 7]. Then individual-level or cluster-level residuals are calculated depending on the type of outcome and the parameter of interest. The residuals from the control and intervention groups are then compared using the standard $t-$test. In the absence of an intervention effect, the residuals are expected to be similar on average between the control and intervention groups. However, if there is an intervention effect, the residuals should differ systemically between the two groups.

1.3.2 Individual-level analysis

As cluster-level analysis may be less efficient in the case of variable sized clusters, more power as well as precision could be obtained by weighting the cluster specific summaries according to the amount of information provided by each cluster. An individual-level analysis, which is essentially a single-stage method, takes into account cluster size and intraclass correlation by performing individual level regression analysis. A wide range of regression models have been proposed in the literature depending on the type of outcome. Two widely accepted regression models are *random effects models* estimated by maximum likelihood methods and *population averaged model* estimated by generalised estimating equations (GEE).

Random effects models take into account between-cluster variability using cluster-level effects which follow a specified probability distribution. The parameters of this distribution are estimated using maximum likelihood methods together with the fixed effect coefficients corresponding to intervention effect and other covariates effects, if any. Depending on the parameter of interest and type of outcome, the most commonly used random effects models are *linear mixed model* (LMM) for quantitative outcomes, *random effects logistic regression model* (RELR) for binary outcomes and *random effects Poisson regression model* for time-to-event outcomes. These regression methods will be explained in the following chapters.

Chapter 2

Missing Data in CRTs

In this chapter, we describe the issue of missing data in CRTs. In Section 2.1, we explain why missing data is a big issue in the analysis of CRTs. Section 2.2 describes Rubin's framework for missingness mechanisms in general and in the context of CRTs with missing outcomes. Finally, in Section 2.3, we review the methods which are most commonly used to handle missing data in CRTs with their pros and cons.

2.1 Is missing data a big issue in CRTs?

Attrition is common in CRTs, leading to missing outcome data that often create a problem in the analysis of such trials. Not only do they cause a loss of information and as a result usually reduce the power of a study, but also they might be a potential

source of bias in the parameter estimates, which itself may lead to statistical tests of the null hypothesis of no intervention effect to be invalid [8, 9]. Handling missing data in CRTs is complicated compared to that of standard trials by the fact that data are clustered in CRTs, that is, the outcomes of individuals within the same cluster are more likely to be similar to each other than those from different clusters, which is usually quantified by the intraclass correlation coefficient. Most of the standard missing data methods like multiple imputation (MI) assume non-clustered data, so do not automatically accommodate this clustering. Ignoring this clustering in general gives biased estimates, as well as having invalid variance estimates [10]. A systematic review was performed by DiazOrdaz *et al.* [11] to see how missing data are handled and reported in CRTs published in 2011. They found that 95 (72%) trials out of 132 trials had missing values either in outcome or in covariates or in both. Only 32 (34%) trials out of 95 reported how they handled missing data. Another recent systematic review by Fiero *et al* [12] on handling of missing data in CRTs found that 80 (93%) trials out of 86 trials reported missing outcome data at the individual-level. The median percent of individuals with missing outcome was 19%, with a range of 0.5 to 90%. Of those trials reporting missing data, only 30 (38%) trials reported how they handled missing outcome data. Despite missing data being very common in CRTs, these two systematic reviews show that handling missing data in CRTs remains suboptimal. One of the key reasons may be that methodological development for dealing with missing data in CRTs has been relatively slow in spite of the increasing popularity of CRTs. Therefore, methods for handling missing data in CRTs need more attention.

The main reasons for having missing outcomes include dropout, withdrawal or lost to follow-up. Withdrawals could occur due to perceived/actual lack of efficacy of intervention, due to condition improving or worsening, and adverse events and therefore no longer wanting to participate in the trial. For example, in a systolic blood pressure (SPB) control trial, younger people may be more likely to withdraw themselves from the study in the sense that they might think high SBP is a problem for older people.

2.2 Missingness mechanism

In statistical analysis, if there are missing values, an assumption must be made about the missingness mechanism, which refers to the relationship between the probability of data being missing and the underlying values of the variables involved in the analysis [13]. The mechanisms which caused the data to be missing can be classified into three broad categories introduced by Rubin [14]. These are *missing completely at random* (MCAR), *missing at random* (MAR), and *missing not at random* (MNAR). Although we do not know the true mechanism for missing values, we assume that there exists a true underlying missingness mechanism, and given a set of variables, we can define what it means for that mechanism to be one of these three categories.

Data are said to be MCAR if the probability of a value being missing is independent of the observed and unobserved data. It implies that causes of missingness are not related to the data. It is generally a very restrictive assumption of missingness and unlikely to hold in many studies. A more realistic assumption of missingness mechanism than MCAR for many studies is MAR where, conditioning on observed data, the probability

of a value being missing is independent of the unobserved data. Data are defined to be MNAR if the probability of a value being missing depends on both observed and unobserved data.

In CRTs, an assumption regarding missing outcomes that is sometimes plausible is that missingness depends on baseline covariates, but conditioning on these baseline covariates, not on the outcome itself. We refer to this as covariate dependent missingness (CDM). This is an example of MAR when baseline covariates are fully observed. A further possibility in CRTs is that whole clusters can be missing. This can occur for example if a cluster withdraw itself from the study. However, in this thesis, we restrict out attention to missing individual outcomes under CDM assumption, and assume that all baseline covariates are fully observed.

2.3 Methods used to handle missing data in CRTs

The impact of missing data on estimation and inference of a parameter of interest depends on the mechanism that caused missing data, the method used to handle missing data, and the choice of statistical methods used for data analysis. The systematic review done by DiazOrdaz *et al.* [11] revealed that 32 trials out of 95 trials explained how they handled missing data. Twenty two of them used a variety of single imputation, namely regression imputation, mean imputation and last observation carried forward (LOCF) for quantitative data and best/worst case for binary data, 8 used multiple imputation

without considering the clustering, and 2 used likelihood-based complete case analysis assuming MAR. We now describe the most commonly used methods that are used in CRTs to handle missing outcome data with their advantages and disadvantages.

2.3.1 Complete records analysis

In complete records analysis (CRA), often referred to as complete case analysis, only individuals with complete data on all variables in the analysis are considered. CRA is widely used because of its simplicity and it is usually the default method of most statistical software packages. Discarding individuals might be a potential source of loss of information, which leads to loss of precision in the parameter estimates. It is well known that CRA is valid when the missing data mechanism is MCAR. In the case of individual-level regression based analysis, CRA is also valid if, conditioning on covariates, missingness is independent of outcome and the outcome model being fitted is correctly specified [13]. Greonwold *et al.* [15] showed that, in the event of missing outcome under MAR for individually randomised trials, CRA with covariate adjustment gives unbiased estimates with coverage close to nominal level.

2.3.2 Single imputation

If a data set contains a large number of variables then discarding incomplete cases corresponding to each variable may result in a very small data set. This is because observed values of a particular variable are deleted when they belong to cases that have missing values for other variables. Instead of discarding incomplete cases, single imputation

imputes a single value for each missing value and creates an artificial complete data set. There are several possible choices for single imputation in the missing data literature, but here we present briefly some of them that are commonly used to the analysis of CRTs. It is important to note that a problem common to all single imputation methods in general is that the subsequent confidence intervals and tests are not valid because no allowance is made for the imputation uncertainty.

2.3.2.1 Last observation carried forward

Last observation carried forward (LOCF) method is usually used in longitudinal studies where repeated measures are taken from each individual at a series of planned follow-up visits. Missing outcome values are replaced with the corresponding individual's last observation, assuming that the missing value for an individual is exactly the same as the previous measurement of that individual. It is usually implausible that an individual's outcome would remain same after withdrawal from the study. The method LOCF has been shown to be invalid in general in non-clustered trials [16], and therefore one cannot expect it to be valid in CRTs either.

2.3.2.2 Mean imputation for continuous outcome

In this case, in general, missing values of a variable are substituted by the mean of the available observed values of that variable. As a result, the mean of the variable remains same but other features (for example, variance, skewness, kurtosis and so forth) of its distribution are changed. Clearly, this method leads to underestimation of the variance

since, for each imputed value, the squared deviation from its mean is zero but the number of observations is increased. However, one can fix estimators of the variance by modifications to account for the mean imputation, for example, by adjusting the degrees of freedom for the imputed values.

Two choices for mean imputation for missing outcomes that have been considered in CRTs are intervention group mean imputation and cluster mean imputation. In the first case, missing outcomes in each intervention group are replaced by the mean calculated using the observed outcomes pooled across clusters of that group. By imputing the intervention group mean for missing outcome values, the variability among the cluster means is reduced [17]. Thus, group mean imputation may give inflated type I error to the null hypothesis. In the latter case, missing outcomes in each cluster are replaced by the mean calculated using the observed outcomes of that cluster. In this case, the imputed cluster means are identical with the observed cluster means.

Cluster mean imputation for missing continuous outcomes has been suggested as a good approach for handling missing outcome data in CRTs by Taljard *et al* [17]. They demonstrated the impact of cluster mean imputation on the validity and power of adjusted $t-$test (describe in Chapter 4) for intervention effect using individual-level outcome data.

2.3.2.3 Regression imputation

In regression imputation, missing outcome values are predicted from the individual's observed, for example, baseline covariates using a model based on observed cases. The fundamental assumption is that missing outcomes can be estimated by the individuals' observed covariate values. In the case of individually randomised trials, this method provides unbiased estimate under MAR [11] but underestimates the standard error like other single imputation methods [8].

2.3.3 Multiple imputation

Multiple imputation (MI), first proposed by Rubin (1987) [18], is a method for filling in the missing values multiple times by simulating from an appropriate model. The aim of imputing multiple times is to allow for the uncertainty associated with the imputed values due to the fact that the imputed values are sampled draws for the missing values instead of the actual values. This uncertainty is taken into account by adding between-imputation variance to the average within-imputation variance.

Multiple imputation method can be summarised in three steps as

1. **Imputation step:** A sequence of T imputed data sets are obtained by replacing each missing value by a set of $T \geq 2$ imputed values that are simulated from an appropriate distribution or model.

2. **Analysis step:** Each of the T data sets are then analysed as a completed data set using the full data analysis method.

3. **Combination step:** The results obtained in the analysis step are combined for inference using Rubin's rules [18].

We now describe Rubin's rules for estimating parameter θ. After analysing T imputed data sets, we have T estimates of θ, denoted here as $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_T$, with their associated variance estimates $\widehat{\text{Var}}(\hat{\theta}_1), \widehat{\text{Var}}(\hat{\theta}_2), \ldots, \widehat{\text{Var}}(\hat{\theta}_T)$, respectively. As described by Rubin [18], the combined estimate of θ can be calculated as

$$\hat{\theta}_{MI} = \frac{1}{T}\sum_{t=1}^{T}\hat{\theta}_t.$$

The variance of this combined estimator has two parts: within-imputation variance and between-imputation variance. The average within-imputation variance, denoted by $\bar{W}_{MI}$, and the between-imputation variance, denoted by B_{MI}, are calculated as, respectively,

$$\bar{W}_{MI} = \frac{1}{T}\sum_{t=1}^{T}\widehat{\text{Var}}(\hat{\theta}_t) \quad \text{and} \quad B_{MI} = \frac{1}{T-1}\sum_{t=1}^{T}\left(\hat{\theta}_t - \hat{\theta}_{\text{MI}}\right)^2.$$

Then the total variability of $\hat{\theta}_{\text{MI}}$, denoted by σ^2_{MI}, is estimated by

$$\hat{\sigma}^2_{\text{MI}} = \bar{W}_{MI} + (1 + T^{-1})B_{MI},$$

where the term $(1 + T^{-1})$ is an adjustment for finite T, the number of imputations.

MI produces valid inferences in general under MAR, provided that the imputation model is correctly specified. One advantage of MI is that it can provide consistent estimates with valid confidence interval with a low number of imputations [19]. However, it is recommended to use a higher number of imputations for a precise estimate by reducing the Monte-Carlo error, leading to a reduction in standard error of the estimate. One important feature of MI is that the imputation model and the analysis model do not have to be the same. However, in order for Rubin's rules to be valid, in general, the imputation model needs to be compatible or congenial with the analysis model in the sense that the imputation model has to contain the analysis model [20]. There may be two possible kinds of uncongenial models. First, the imputation model is simpler than the analysis model. For example, the imputation model is linear, but the analysis model includes interactions and non-linearities. In this situation, Rubin's variance formula is invalid and arguably more importantly the parameter estimates from the analysis model are not consistent. Second, the imputation model is richer than the analysis model. For example, the imputation model uses auxiliary variables that are not involved in the analysis model. In this case, the variability of the MI estimator may be overestimated by Rubin's variance formula.

There are at least four different types of MI that have been used in CRTs [11]. These are *standard* MI, also known as *single-level* MI, which ignores clustering in the imputation model, *fixed effects* MI which includes a fixed effect for each cluster in the imputation model, *random effects* MI where clustering is taken into account through a random effect for each cluster in the imputation model, and *within-cluster* MI where standard MI is applied within each cluster. From now, we refer to random effects MI as multilevel multiple imputation (MMI).

In the case of missing continuous outcome in CRTs, Andridge [10] showed that the true MI variance of group means are underestimated by single-level MI, and are overestimated by fixed effects MI. She also showed that MMI is the best among these three methods and recommended its use for practitioners. DiazOrdaz *et al.* [21] showed that for bivariate outcomes MMI gives coverage rate close to nominal level, whereas single-level MI gives low coverage and fixed effects MI gives overcoverage.

Gomes *et al.* [22] investigated the performance of MMI in cost-effectiveness analysis (CEA) compared to single-level MI and CRA. In their study, missingness was in both cost and outcome variables, but covariates were fully observed. They assumed that the error terms of the imputation models for costs and outcomes follow a bivariate normal distribution; and, in addition, the cluster-specific random effects for costs and outcomes follow a bivariate normal distribution. Different scenarios of missingness were considered under MCAR, MAR and MNAR. They also considered both cluster-level and individual-level covariates that predicted missing values in cost and outcome variables. They concluded that the point estimates of cost-effectiveness and standard errors using MMI were close to those estimates using fully observed data, under MAR and MNAR, compared to single-level MI and CRA. However, it is not clear whether the CRA was adjusted for the fully observed covariates. Under MCAR, the estimates of cost-effectiveness for each approach were similar to those from the fully observed data. DiazOrdaz *et al.* [23] also presented MMI as a better approach for handling missing values in CEAs compared to single-level MI and CRA. The study was illustrated with CEAs that use data from CRTs with missingness in both cost and outcome variables. They considered bivariate normal distribution to represents random cluster effects for

cost and outcome variables. The incremental cost, incremental quality-adjusted life-years and incremental net benefit were estimated for each approach and compared. The results showed that the estimates obtained using CRA are biased.

Taljaard *et al.* [17] examined the performance of MI for missing continuous outcomes in CRTs in a simple setup where there is no covariates except intervention indicator. They used standard MI and MMI. They also considered the Approximate Bayesian Bootstrap (ABB) procedure, proposed by Rubin and Schenker [24], as non-parametric MI. In ABB, sampling from the posterior predictive distribution of missing outcomes is approximated by first generating a set of plausible contributors drawn with replacement from the observed data, and then imputed values are drawn with replacement from the possible contributors. Two types of ABB in CRTs investigated were pooled ABB and within-cluster ABB, where the set of possible contributors are sampled from all observed values across the clusters in each intervention group or from observed values in the same cluster, respectively. They showed that none of these four MI procedures tend to yield better power compared to the power of adjusted $t-$test using no imputation or cluster mean imputation under MCAR. In the case of missing outcome under MAR in non-clustered trials, Groenwold *et al.* [15] showed that CRA with covariate adjustment and MI give similar estimates so long as the same functional form of the same set of predictors of missingness are used.

In the case of missing binary outcomes in CRTs, Ma *et al.* [25] examined within-cluster MI, fixed effects MI and MMI under CDM mechanism in CRTs. They showed that all these strategies give similar performance in terms of bias with low percentages of missing data or with small value of ICC. With high percentage of missing data, they

concluded that within-cluster MI underestimated the variance of the intervention effect which result in inflated Type I error rate. In two subsequent studies, Ma *et al.* [26, 27] compared the performance of GEE and RELR with missing binary outcomes using standard MI and within-cluster MI. Results showed that GEE performs well when using standard MI and the variance inflation factor (VIF) is less than 3; and using within-cluster MI when VIF $\geq$ 3 and cluster size is at least 50. Ma *et al.* [27] concluded that RELR does not perform well using either standard MI or within-cluster MI. Caille *et al.* [28] compared different MI strategies through a simulation study for handing missing binary outcomes in CRTs assuming CDM. They showed that MMI with RELR and single-level MI with standard logistic regression give better inference for intervention effect compared to CRA in terms of bias, efficiency and coverage.

2.4 Outline of the thesis

In this thesis, we will review the literature in greater detail and evaluate the methods for each of the three outcomes types in CRTs, where outcomes are missing under CDM mechanism. Part-II consists of three chapters, and deals with continuous outcomes. Part-III has two chapters and deals with binary outcomes. Part-IV deals with time-to-event outcomes. Part-V summarises the findings of this thesis and discusses possible extensions and future work.

Part II

Continuous Outcomes

Chapter 3

Review of Analysis Methods with Full Data

In this chapter, we discuss the terminology and define the necessary notation to be used in the next two chapters for continuous outcomes, and describe the methods for handling missing continuous outcomes in CRTs. Section 3.1 defines the notation for the variables involved. In Section 3.2 and Section 3.3, we describe standard $t-$test and adjusted $t-$test for testing intervention effect in CRTs with full data. Section 3.4 explains LMM as the individual-level analysis. In Section 3.5, we conclude the chapter by outlining what we will investigate in the next two chapters for continuous outcomes.

3.1 Notations

Consider a two arm CRT. Let Y_{ijl} be a continuous outcome for the lth ($l = 1, 2, \ldots, m_{ij}$) individual in the jth ($j = 1, 2, \ldots, k_i$) cluster of the intervention group i ($i = 0, 1$), where $i = 0$ corresponds to control group, $i = 1$ corresponds to active intervention group.

Let each outcome Y_{ijl} be generated by a linear mixed model (LMM)

$$Y_{ijl} = \mu_i + \delta_{ij} + \epsilon_{ijl}, \tag{3.1}$$

where μ_i is the mean of the ith intervention group and $\delta_{ij} \sim N(0, \sigma_b^2)$ independently of $\epsilon_{ijl} \sim N(0, \sigma_w^2)$. Then $\mathrm{E}_{jl}(Y_{ijl}) = \mu_i$, $\mathrm{Var}(Y_{ijl}) = \sigma_b^2 + \sigma_w^2 (= \sigma^2)$, and $\mathrm{Cov}(Y_{ijl}, Y_{ijs}) = \sigma_b^2$ for $l \neq s$, where σ_b^2 and σ_w^2 denote the between-cluster variability and the within-cluster variability, respectively, and σ^2 denotes the total variance. The quantity $\mu_1 - \mu_0$ represents the size of the intervention effect. Note that model (3.1) does not contain any baseline covariates. We will consider baseline covariates in the following chapters. In the following two sections we explain the standard $t-$test and adjusted $t-$test in the absence of missing data.

3.2 Standard $t-$test

In the cluster-level analysis methods, the standard $t-$test for two independent samples (here referred to as cluster-level $t-$test) is the most commonly used method to compare the means of the control group and intervention group. Suppose $\bar{Y}_{ij}$ is the mean of

outcome Y_{ijl} in the (ij)th cluster, defined by,

$$\bar{Y}_{ij} = \frac{1}{m_{ij}} \sum_{l=1}^{m_{ij}} Y_{ijl}.$$

Assuming $\bar{Y}_{ij}$ follows normal distribution with mean $\mu_i (i = 0, 1)$ and common variance σ_c^2, a test statistic for the null hypothesis of no intervention effect, symbolically $H_0 : \mu_1 = \mu_0$, is given by

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_0}{S\sqrt{\frac{1}{k_0} + \frac{1}{k_1}}} \sim t_{(k_0+k_1-2)}, \tag{3.2}$$

where $\hat{\mu}_i$ is the estimated mean of the ith intervention group and S^2 is the pooled estimate of the common variance σ_c^2 computed as, respectively,

$$\hat{\mu}_i = \frac{1}{k_i} \sum_{j=1}^{k_i} \bar{Y}_{ij} \qquad \text{and} \qquad S^2 = \frac{\sum_{i=0}^{1} \sum_{j=1}^{k_i} \left(\bar{Y}_{ij} - \hat{\mu}_i\right)^2}{k_0 + k_1 - 2}.$$

Note that, in the calculation of $\hat{\mu}_i$, equal weights are given to the cluster means ignoring the variation in cluster sizes. The validity of this test depends on the underlying assumption $\bar{Y}_{ij} \sim N\left(\mu_i, \sigma_c^2\right)$ for $i \in \{0, 1\}$, that is, cluster means are normally distributed with mean depending on i (intervention group index) and with common variance across the intervention groups. The normality assumption is guaranteed by the central limit theorem if the cluster sizes are sufficiently large. Also it has been shown by simulation that the $t-$test is robust in terms of deviations from normality when the intervention groups have the equal number of clusters, even for small number of cluster [5]. Under the null hypotheses, the assumption that the variance is constant across the intervention groups is guaranteed by the random allocation of clusters between the intervention groups [5].

3.3 Adjusted $t-$test

The adjusted $t-$test, proposed by Donner and Klar (2000) [2], is an alternative approach to test the intervention effect for quantitative outcomes using individual-level data. This test is a simple extension of standard $t-$test.

Let $M_i = \sum_{j=1}^{k_i} m_{ij}$ be the total number of individuals in the ith intervention group. Also let $M = \sum_{i=0}^{1} M_i$ and $K = \sum_{i=0}^{1} k_i$ be the total number of clusters, respectively, in the study. Then assuming Y_{ijl} is normally distributed with mean μ_i and variance $\sigma^2 = \sigma_b^2 + \sigma_w^2$, a test for $H_0 : \mu_1 = \mu_0$ based on standard $t-$test adjusted for intraclass correlation is given by [2]

$$t_A = \frac{\tilde{\mu}_1 - \tilde{\mu}_0}{\widehat{\mathrm{SE}}(\tilde{\mu}_1 - \tilde{\mu}_0)} \sim t_{K-2} \tag{3.3}$$

where

$$\tilde{\mu}_i = \frac{1}{M_i} \sum_{j=1}^{k_i} \sum_{l=1}^{m_{ij}} Y_{ijl} = \frac{\sum_{j=1}^{k_i} m_{ij} \bar{Y}_{ij}}{M_i}, \quad i = 0, 1$$

is the estimated mean of the ith intervention group, which is calculated by taking the cluster size as weight for each cluster mean, and

$$\widehat{\mathrm{SE}}(\tilde{\mu}_1 - \tilde{\mu}_0) = \sqrt{S_P^2 \left(\frac{\widehat{\mathrm{VIF}}_1}{M_1} + \frac{\widehat{\mathrm{VIF}}_0}{M_0} \right)}$$

is the estimated standard error of $(\tilde{\mu}_1 - \tilde{\mu}_0)$, where $S_P^2 = S_w^2 + S_b^2$ is the pooled estimate of total the variance $\sigma^2 = \sigma_b^2 + \sigma_w^2$ and $\widehat{\text{VIF}}_i = 1 + (A_i - 1)\hat{\rho}$ is the variance inflation factor for intervention group i $(i = 0, 1)$ [29] with $A_i = \sum_{j=1}^{k_i} m_{ij}^2 / M_i$ and

$$\hat{\rho} = \frac{\text{MSC} - \text{MSW}}{\text{MSC} + (m_0 - 1)\text{MSW}},$$

where MSW and MSC are the within-cluster mean square error and between-cluster mean square error, respectively, and $m_0 = \left(M - \sum_{i=0}^{1} A_i\right)/(K-2)$. An equivalent expression for $\hat{\rho}$ can be written as $\hat{\rho} = S_b^2/(S_w^2 + S_b^2)$, where $S_b^2 = (\text{MSC} - \text{MSW})/m_0$ and $S_w^2 = \text{MSW}$ are the analysis of variance (ANOVA) estimates of σ_b^2 and σ_w^2, respectively, and

$$\text{MSW} = \frac{1}{M-K} \sum_{i=0}^{1} \sum_{j=1}^{k_i} \sum_{l=1}^{m_{ij}} \left(Y_{ijl} - \bar{Y}_{ij}\right)^2$$

and

$$\text{MSC} = \frac{1}{K-2} \sum_{i=0}^{1} \sum_{j=1}^{k_i} m_{ij} \left(\bar{Y}_{ij} - \tilde{\mu}_i\right)^2.$$

3.4 Linear Mixed Model

The linear mixed model (LMM) takes into account between-cluster variability using cluster-level effects which are assumed to follow a specified probability distribution. The parameters of that distribution are estimated using maximum likelihood methods together with intervention effect and other covariates effects. However, the variances of the fixed effect parameters estimates, which are calculated based on their asymptotic distributions, are known to be underestimated for small sample size [30]. In practice, for testing hypotheses about fixed-effects parameters, the resulting downward bias is often

handled by using approximate $t-$statistic and $F-$statistic [31]. An approximate test can be obtained by approximating the distribution of the test-statistics by a $t-$distribution. Satterthwaite [32] proposed an approximation, known as Satterthwaite's approximation, to calculate the degrees of freedom of the $t-$distribution. For testing hypotheses of the form $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$, where $\boldsymbol{\beta}$ is a vector of fixed-effects parameters and $\mathbf{L}$ is any known matrix, Kenward and Roger [30] suggested a scaled Wald statistic as well as an F approximation of its sampling distribution that performs well for small sample size. The suggested statistic uses an adjusted estimate of the variance-covariance matrix that has minimum bias due to small sample size. The numerator degrees of freedom of the approximate $F-$distribution equals rank($\mathbf{L}$) and the denominator degrees of freedom is calculated via a Satterthwaite-type approximation [31]. As far as we are aware no study has been done to use these two approximations in CRTs. Both these approximation are applicable for LMM and related multivariate normally based models [33]. Kenward-Roger's approximation essentially recovers Satterthwaite's approximation when there is only one fixed effect in the model [30].

3.5 Summary

In Chapter 4, we investigate (a) the impact of cluster mean imputation for missing continuous outcome values on the variance components estimates, and (b) the impact of small number of clusters in each intervention groups on the validity of LMM with full data, CRA and cluster mean imputation.

Chapter 5 is a published paper that investigates the performance of cluster-level analyses and individual-level analysis under CDM in continuous outcomes in terms of bias, average estimated standard error and coverage rate.

Chapter 4

Cluster Mean Imputation

This chapter investigates the impact of cluster mean imputation for missing continuous outcomes on the variance components estimates. In addition, it investigates the impact of small number of clusters in each intervention group on the validity of LMM analysis with full data, CRA and cluster mean imputation. Section 4.1 explains cluster mean imputation method for handling missing outcomes in CRTs. Section 4.2 describes two different examples of MCAR and one example of MAR for missing outcomes in the context of CRTs. In Section 4.3, we investigate analytically the validity of the ANOVA estimators of the variance components with cluster mean imputation for missing continuous outcomes. Section 4.4 describes a simulation study and presents the results to support the derived analytical results in Section 4.3. We compare analytically cluster-

level $t-$test, adjusted $t-$test and LMM under balanced CRT in Section 4.5. In Section 4.6, we conduct another simulation study to investigate the impact of CRA, cluster mean imputation for missing outcomes on the validity and power of cluster-level $t-$test, adjusted $t-$test and LMM under MCAR and MAR. Section 4.7 concludes this chapter with some discussion. Note that in this chapter we do not consider any baseline covariate except the intervention indicator.

4.1 Cluster mean imputation

Suppose the outcome variable is partially observed. In cluster mean imputation, missing continuous outcomes in each cluster are replaced by the observed mean calculated using the observed values of that cluster. Define an indicator variable R_{ijl} such that

$$R_{ijl} = \begin{cases} 1, & \text{if } Y_{ijl} \text{ is observed} \\ 0, & \text{if } Y_{ijl} \text{ is missing.} \end{cases} \tag{4.1}$$

Let $\bar{Y}_{ij}^{\text{obs}}$ be the observed mean of the (ij)th cluster calculated as

$$\bar{Y}_{ij}^{\text{obs}} = \frac{1}{W_{ij}} \sum_{l=1}^{m_{ij}} R_{ijl} Y_{ijl},$$

where $W_{ij} = \sum_{l=1}^{m_{ij}} R_{ijl}$ is the number of observed outcomes in the (ij)th cluster. The observed cluster mean imputation results in a complete dataset $\mathbf{Y}^* = (Y^*_{ijl})$ such that

$$Y^*_{ijl} = \begin{cases} Y_{ijl}, & R_{ijl} = 1 \\ \bar{Y}^{\text{obs}}_{ij}, & R_{ijl} = 0, \end{cases} \tag{4.2}$$

where * refers to the completed data through cluster mean imputation. Then the imputed cluster means $(\bar{Y}^*_{ij})$ are identical with the observed cluster means $(\bar{Y}^{\text{obs}}_{ij})$ due to cluster mean imputation.

Taljaard *et al.* [17] investigated Type I error and power of the adjusted $t-$test for intervention effect, considering balanced CRT, using cluster mean imputation under MCAR. They found that cluster mean imputation yields acceptable Type I error and suggested it may be a good approach for missing outcome data in CTRs. However, it might give lower power compared to other imputation procedures when the cluster sizes are small and cluster follow-up rates are varied highly. Moreover, they did not consider the consequence of this imputation to the parameter estimates of within-cluster variance and between-imputation variance. One might have interest to the estimates of the variance components as well. Furthermore, they did not mention any advantage of using adjusted $t-$test over cluster-level $t-$test and LMM using CRA in the case of balanced CRTs with no covariates except intervention indicator.

4.2 Missingness mechanism in CRTs

In CRTs, the missingness of an individual's outcome may depends on the characteristics of the individual, intervention and cluster or the other individuals of the same cluster. Let $\mathbf{Y}_{ij} = (Y_{ij1}, Y_{ij2}, \cdots, Y_{ijm_{ij}})'$ be the vector of m_{ij} values that were intended to be observed in the jth cluster of the ith intervention group. Assume that $\mathbf{Y}_{ij}$ is partially observed. The missingness mechanism for an outcome Y_{ijl} can be represented by the conditional probability $P(R_{ijl} = 0|\mathbf{Y}_{ij})$. Then

- Data are said to be MCAR if

$$P(R_{ijl} = 0|\mathbf{Y}_{ij}) = P(R_{ijl} = 0) \quad \forall i, j, l. \tag{4.3}$$

  For example, in a trial of a weight reduction program, a measurement may be missing due to running out of batteries in the weighing scale. In this case, the missingness mechanism is plausibly independent of both the intervention and the outcomes.

  Let π_{ijl} be the probability that individual l in the jth cluster of the ith intervention group will have their outcome observed. Then

$$R_{ijl} \sim \text{Bernoulli}(\pi_{ijl}).$$

  In CRTs, one possibility of MCAR could be that each cluster has the same follow-up rate (here referred to as MCAR1) regardless of the intervention group. Another possibility could be that cluster follow-up rates vary randomly and independently

of the intervention groups and $\mathbf{Y}_{ij}$ (here referred to as MCAR2). If π $(0 < \pi < 1)$ is the constant follow-up rate in each cluster and if π_{ij} $(0 < \pi_{ij} < 1)$ is the follow-up rate for the (ij)th cluster which varies randomly and independently of the intervention group and $\mathbf{Y}_{ij}$, then data are said to be

$$\text{MCAR1 if } \pi_{ijl} = \pi \quad \forall i, j, l \tag{4.4}$$

and

$$\text{MCAR2 if } \pi_{ijl} = \pi_{ij} \quad \forall i, j, l \tag{4.5}$$

- Data are said to be MAR if

$$P(R_{ijl} = 0|\mathbf{Y}_{ij}) = P(R_{ijl} = 0|\mathbf{Y}_{ij}^{\text{obs}}) \quad \forall i, j, l \tag{4.6}$$

where $\mathbf{Y}_{ij}^{\text{obs}}$ denotes the vector of observed outcomes in the jth cluster of the ith intervention group. Consider a trial of systolic blood pressure (SBP) control where a measurement on SBP may be missing depending on whether an individual receives intervention in a sense that individuals in the intervention group are more likely to have observed SBP measurement. One of the reasons may be that individuals in the intervention group are usually enthusiastic to check their SBP level. Therefore, after conditioning on intervention group, missingness in SBP measurement is independent of the value of SBP measurement. An example of an MAR mechanism could be

$$P(R_{ijl} = 0|\mathbf{Y}_{ij}) = 1 - \pi_i \quad \forall i, j, l \tag{4.7}$$

where $\pi_i (0 < \pi_i < 1)$ is the constant follow-up rate in the ith intervention group. In this example, since missingness only depends on the intervention groups but not on the outcome, it is essentially an example of MCAR within each intervention group.

4.3 Validity of S_w^2 and S_b^2 with cluster mean imputation

This section investigates the unbiasedness of the ANOVA estimators of variance components, within-cluster variance (σ_w^2) and between-cluster variance (σ_b^2), with cluster mean imputation when outcomes are missing under MCAR or MAR mechanism given in equations (4.4) and (4.7), respectively. For simplicity, we restrict our attention to balanced CRTs without any covariates except intervention indicator.

Considering the model as given in equation (3.1), the mean of (ij)th cluster can be written as

$$\begin{aligned} \bar{Y}_{ij} &= \frac{1}{m}\sum_{l=1}^{m} Y_{ijl} \\ &= \mu_i + \delta_{ij} + \frac{1}{m}\sum_{l=1}^{m} \epsilon_{ijl} \\ &= \mu_i + \delta_{ij} + \bar{\epsilon}_{ij}. \end{aligned}$$

and the mean of the ith intervention group is

$$\bar{Y}_i = \frac{1}{k}\sum_{j=1}^{k} \bar{Y}_{ij}.$$

The variance of $\bar{Y}_{ij}$ and $\bar{Y}_i$ are then given by

$$\text{Var}(\bar{Y}_{ij}) = \sigma_b^2 + \frac{\sigma_w^2}{m} \quad \text{and} \quad \text{Var}\left(\bar{Y}_i\right) = \frac{1}{k}\left(\sigma_b^2 + \frac{\sigma_w^2}{m}\right).$$

Suppose the outcome Y_{ijl} is partially observed. Then the observed mean of the (ij)th cluster can be calculated as

$$\begin{aligned} \bar{Y}_{ij}^{\text{obs}} &= \frac{1}{W_{ij}} \sum_{l=1}^{m} R_{ijl} Y_{ijl} \\ &= \mu_i + \delta_{ij} + \frac{1}{W_{ij}} \sum_{l=1}^{m} R_{ijl} \epsilon_{ijl} \\ &= \mu_i + \delta_{ij} + \bar{\epsilon}_{ij}^{\text{obs}}, \end{aligned}$$

where $W_{ij} = \sum_{l=1}^{m} R_{ijl}$ is the number of observed outcomes in the (ij)th cluster. The observed mean of the ith intervention group is then calculated as

$$\bar{Y}_i^{\text{obs}} = \frac{1}{k} \sum_{j=1}^{k} \bar{Y}_{ij}^{\text{obs}}.$$

The conditional mean and variance of $\bar{Y}_{ij}^{\text{obs}}$, given W_{ij}, is

$$\text{E}_j(\bar{Y}_{ij}^{\text{obs}} | W_{ij}) = \mu_i \quad \text{for} \quad i \in \{0, 1\} \tag{4.8}$$

$$\text{Var}_j(\bar{Y}_{ij}^{\text{obs}} | W_{ij}) = \sigma_b^2 + \frac{\sigma_w^2}{W_{ij}} \quad \text{for} \quad i \in \{0, 1\}. \tag{4.9}$$

The unconditional variance of $\bar{Y}_{ij}^{\text{obs}}$ can be found by averaging over W_{ij} as

$$\text{Var}_j(\bar{Y}_{ij}^{\text{obs}}) = \text{E}_j\big[\text{Var}_j(\bar{Y}_{ij}^{\text{obs}} | W_{ij})\big] + \text{Var}_j\big[\text{E}_j(\bar{Y}_{ij}^{\text{obs}} | W_{ij})\big], \quad i \in \{0, 1\}. \tag{4.10}$$

The second term of the right hand side of (4.10) becomes zero since, from equation (4.8), $\mathrm{E}_j\left(\bar{Y}_{ij}^{\mathrm{obs}}|W_{ij}\right)$ is constant for $i \in \{0,1\}$. Then, plugging the result in equation (4.9) into equation (4.10), we get

$$\begin{aligned}\mathrm{Var}_j(\bar{Y}_{ij}^{\mathrm{obs}}) &= \mathrm{E}_j\left(\sigma_b^2 + \frac{\sigma_w^2}{W_{ij}}\right) \\ &\approx \sigma_b^2 + \frac{\sigma_w^2}{\mathrm{E}_j(W_{ij})}, i \in \{0,1\}, \end{aligned} \tag{4.11}$$

using delta method which will be valid if $\mathrm{Var}_j(W_{ij})$ is small.

Under MCAR1, defined in (4.4), $W_{ij} \sim \mathrm{Bin}(m,\pi)$ and, hence, $\mathrm{E}(W_{ij}) = m\pi, \forall i,j$. Therefore, the variance of $\bar{Y}_{ij}^{\mathrm{obs}}$ can be written as

$$\mathrm{Var}_j\left(\bar{Y}_{ij}^{\mathrm{obs}}\right)_{\mathrm{MCAR1}} \approx \sigma_b^2 + \frac{\sigma_w^2}{m\pi},\ i \in \{0,1\} \tag{4.12}$$

Under MAR, defined in (4.7), $W_{ij} \sim \mathrm{Bin}(m,\pi_i)$ and, hence, $\mathrm{E}(W_{ij}) = m\pi_i, \forall i,j$. Then the variance of $\bar{Y}_{ij}^{\mathrm{obs}}$ can be written as

$$\mathrm{Var}_j\left(\bar{Y}_{ij}^{\mathrm{obs}}\right)_{\mathrm{MAR}} \approx \sigma_b^2 + \frac{\sigma_w^2}{m\pi_i},\ i \in \{0,1\} \tag{4.13}$$

In the case of balanced CRTs, the ANOVA estimators of σ_w^2 and σ_b^2 are, respectively, given by

$$S_w^2 = \mathrm{MSW} \qquad \text{and} \qquad S_b^2 = (\mathrm{MSC} - \mathrm{MSW})/m,$$

where MSW and MSC are the within-cluster mean square error and between-cluster mean square error, respectively, and can be written as

$$\text{MSW} = \frac{1}{2k(m-1)} \sum_{i=0}^{1} \sum_{j=1}^{k} \sum_{l=1}^{m} \left(Y_{ijl} - \bar{Y}_{ij}\right)^2$$

and

$$\text{MSC} = \frac{m}{2(k-1)} \sum_{i=0}^{1} \sum_{j=1}^{k} \left(\bar{Y}_{ij} - \hat{\mu}_i\right)^2,$$

where $\hat{\mu}_i$, the mean of the cluster means in the ith intervention group, is an estimate of μ_i, the true mean of ith intervention group. With observed cluster mean imputation for missing outcomes, the MSW can be rewritten as

$$\begin{aligned}
\text{MSW} &= \frac{1}{2k(m-1)} \sum_{i=0}^{1} \sum_{j=1}^{k} \sum_{l=1}^{m} \left(Y^*_{ijl} - \bar{Y}^*_{ij}\right)^2 \\
&= \frac{1}{2k(m-1)} \sum_{i=0}^{1} \sum_{j=1}^{k} \sum_{l=1}^{m} \left[R_{ijl} \left(Y_{ijl} - \bar{Y}^*_{ij}\right)^2 + (1 - R_{ijl}) \left(\bar{Y}^{\text{obs}}_{ij} - \bar{Y}^*_{ij}\right)^2 \right] \\
&= \frac{1}{2k(m-1)} \sum_{i=0}^{1} \sum_{j=1}^{k} \sum_{l=1}^{m} R_{ijl} \left(Y_{ijl} - \bar{Y}^{\text{obs}}_{ij}\right)^2,
\end{aligned}$$

since the imputed cluster means ($\bar{Y}^*_{ij}$) are identical with the observed cluster means ($\bar{Y}^{\text{obs}}_{ij}$) due to cluster mean imputation. Then

$$\text{E}\left(\text{MSW}\right) = \frac{1}{2k(m-1)} \sum_{i=0}^{1} \sum_{j=1}^{k} \text{E}\left[\sum_{l=1}^{m} R_{ijl} \left(Y_{ijl} - \bar{Y}^{\text{obs}}_{ij}\right)^2 \right]. \tag{4.14}$$

The term $R_{ijl}\big(Y_{ijl}-\bar{Y}_{ij}^{\text{obs}}\big)^2$ is non-zero for W_{ij} observed outcomes and zero for remaining $(m-W_{ij})$ missing outcomes. Therefore, the expectation of $\sum_{l=1}^{m} R_{ijl}\big(Y_{ijl}-\bar{Y}_{ij}^{\text{obs}}\big)^2$ depends on W_{ij}, which is a variable. Conditioning on W_{ij}, we can write

$$\begin{aligned}\mathrm{E}\Big[\sum_{l=1}^{m} R_{ijl}\left(Y_{ijl}-\bar{Y}_{ij}^{\text{obs}}\right)^2\Big] &= \mathrm{E}\Big[\mathrm{E}\Big(\sum_{l=1}^{m} R_{ijl}\big(Y_{ijl}-\bar{Y}_{ij}^{\text{obs}}\big)^2\Big|W_{ij}\Big)\Big]\\ &= \mathrm{E}\Big[\mathrm{E}\Big(\sum_{l=1}^{m} R_{ijl}\big(\epsilon_{ijl}-\bar{\epsilon}_{ij}^{\text{obs}}\big)^2\Big|W_{ij}\Big)\Big].\end{aligned}$$

For a given (ij)th cluster, $\sum_{l=1}^{m} R_{ijl}\big(\epsilon_{ijl}-\bar{\epsilon}_{ij}^{\text{obs}}\big)^2$ is the sum of W_{ij} squared deviations of ϵ_{ijl} from its observed mean $\bar{\epsilon}_{ij}^{\text{obs}}$. Therefore, $\mathrm{E}\Big(\sum_{l=1}^{m} R_{ijl}\big(\epsilon_{ijl}-\bar{\epsilon}_{ij}^{\text{obs}}\big)^2\Big|W_{ij}\Big) = (W_{ij}-1)\,\sigma_w^2$, where σ_w^2 is the variance of ϵ_{ijl}. Finally, we get

$$\mathrm{E}\Big[\sum_{l=1}^{m} R_{ijl}\big(Y_{ijl}-\bar{Y}_{ij}^{\text{obs}}\big)^2\Big] = \mathrm{E}(W_{ij}-1)\sigma_w^2.$$

Plugging this result into equation (4.14), we get

$$\mathrm{E}(\text{MSW}) = \frac{1}{2k(m-1)}\sum_{i=0}^{1}\sum_{j=1}^{k}\mathrm{E}(W_{ij}-1)\sigma_w^2. \tag{4.15}$$

Now with cluster mean imputation, the MSC can be written as

$$\text{MSC} = \frac{m}{2(k-1)}\sum_{i=0}^{1}\sum_{j=1}^{k}\big(\bar{Y}_{ij}^{*}-\hat{\mu}_i^{*}\big)^2,$$

where $\hat{\mu}_i^{*}$ is the mean of imputed cluster means $\bar{Y}_{ij}^{*}$ in the ith intervention group. Then

$$\mathrm{E}\,(\text{MSC}) = \frac{m}{2(k-1)}\sum_{i=0}^{1}\sum_{j=1}^{k}\mathrm{E}\,\big(\bar{Y}_{ij}^{*}-\hat{\mu}_i^{*}\big)^2. \tag{4.16}$$

We can write

$$\begin{aligned}
\bar{Y}_{ij}^{*} - \hat{\mu}_{i}^{*} &= \bar{Y}_{ij}^{*} - \frac{1}{k}\sum_{j=1}^{k}\bar{Y}_{ij}^{*} \\
&= \left(1 - \frac{1}{k}\right)\bar{Y}_{ij}^{*} - \frac{1}{k}\sum_{\substack{s\neq j \\ s=1}}^{k}\bar{Y}_{is}^{*} \\
&= \left(1 - \frac{1}{k}\right)\bar{Y}_{ij}^{\text{obs}} - \frac{1}{k}\sum_{\substack{s\neq j \\ s=1}}^{k}\bar{Y}_{is}^{\text{obs}}, \qquad (4.17)
\end{aligned}$$

since the imputed cluster means ($\bar{Y}_{ij}^{*}$) are identical with the observed cluster means ($\bar{Y}_{ij}^{\text{obs}}$) due to cluster mean imputation. Then

$$\begin{aligned}
\mathrm{E}\left(\bar{Y}_{ij}^{*} - \hat{\mu}_{i}^{*}\right)^{2} &= \mathrm{Var}\left(\bar{Y}_{ij}^{*} - \hat{\mu}_{i}^{*}\right), \text{ since } \mathrm{E}\left(\bar{Y}_{ij}^{*} - \hat{\mu}_{i}^{*}\right) = 0 \\
&= \left(1 - \frac{1}{k}\right)^{2}\mathrm{Var}\left(\bar{Y}_{ij}^{\text{obs}}\right) + \frac{k-1}{k^{2}}\mathrm{Var}\left(\bar{Y}_{ij}^{\text{obs}}\right) \\
&= \left(1 - \frac{1}{k}\right)\mathrm{Var}\left(\bar{Y}_{ij}^{\text{obs}}\right) \qquad (4.18)
\end{aligned}$$

- **Case I: MCAR1 missingness mechanism**

  Under MCAR1, defined in (4.4), $W_{ij} \sim \text{Bin}(m, \pi)$ and, hence, $\mathrm{E}(W_{ij}) = m\pi, \forall i, j$. Therefore, from (4.15), we have

$$\mathrm{E}\left(S_{w}^{2}\right) = \frac{m\pi - 1}{m - 1}\sigma_{w}^{2} \neq \sigma_{w}^{2}.$$

  From (4.18), we can write, using (4.12),

$$\mathrm{E}\left(\bar{Y}_{ij}^{*} - \hat{\mu}_{i}^{*}\right)^{2} = \left(1 - \frac{1}{k}\right)\left(\sigma_{b}^{2} + \frac{\sigma_{w}^{2}}{m\pi}\right),$$

and plugging this result into (4.16), we get

$$\begin{aligned} \mathrm{E}\left(\mathrm{MSC}\right) &= \frac{2mk}{2(k-1)}\left(1-\frac{1}{k}\right)\left(\sigma_b^2+\frac{\sigma_w^2}{m\pi}\right) \\ &= m\sigma_b^2+\frac{1}{\pi}\sigma_w^2. \end{aligned}$$

Then

$$\begin{aligned} \mathrm{E}\left(S_b^2\right) &= \frac{1}{m}\Big[\mathrm{E}\left(\mathrm{MSC}\right)-\mathrm{E}\left(\mathrm{MSW}\right)\Big] \\ &= \frac{1}{m}\left(m\sigma_b^2+\frac{1}{\pi}\sigma_w^2-\frac{m\pi-1}{m-1}\sigma_w^2\right) \\ &= \sigma_b^2+\left(\frac{1}{\pi}-\frac{m\pi-1}{m-1}\right)\frac{\sigma_w^2}{m} \neq \sigma_b^2. \end{aligned}$$

Hence, S_w^2 and S_b^2 are biased estimators for σ_w^2 and σ_b^2, respectively, with cluster mean imputation under MCAR1. Since for $m>0$ and $0<\pi<1$,

$$\frac{m\pi-1}{m-1}<1 \qquad \text{and} \qquad \frac{1}{\pi}-\frac{m\pi-1}{m-1}>0.$$

Hence, S_w^2 is downward biased for σ_w^2, whereas S_b^2 is upward biased for σ_b^2. Also since

$$\frac{m\pi-1}{m-1}=\frac{\pi-\frac{1}{m}}{1-\frac{1}{m}}\to\pi \quad \text{as} \quad m\to\infty$$

and

$$\frac{1}{m}\left(\frac{1}{\pi}-\frac{m\pi-1}{m-1}\right)=\left(\frac{1}{m\pi}-\frac{\pi-\frac{1}{m}}{m-1}\right)\to 0 \quad \text{as} \quad m\to\infty,$$

we deduce $\mathrm{E}\left(S_w^2\right)\approx\pi\sigma_w^2$ and $\mathrm{E}\left(S_b^2\right)\approx\sigma_b^2$ for large m. Therefore, with large cluster size, σ_w^2 is underestimated by a factor approximately equal to the constant follow-up rate π and S_b^2 is approximately unbiased for σ_b^2.

- **Case II: MAR missingness mechanism**

  Under MAR mechanism, defined in (4.7), $W_{ij} \sim \text{Bin}(m, \pi_i)$ and, hence, $\text{E}(W_{ij}) = m\pi_i, \forall i, j$. From equation (4.15), we can write

$$\begin{aligned}
\text{E}\left(\text{MSW}\right) &= \frac{\sigma_w^2}{2k(m-1)} \sum_{j=1}^{k} \Big[\text{E}\left(W_{0j} - 1\right) + \text{E}\left(W_{1j} - 1\right)\Big] \\
&= \frac{\sigma_w^2}{2k(m-1)} \sum_{j=1}^{k} \Big[m\pi_0 + m\pi_1 - 2\Big] \\
&= \frac{m\left(\pi_0 + \pi_1\right) - 2}{2(m-1)} \sigma_w^2.
\end{aligned}$$

  Hence

$$\text{E}\left(S_w^2\right) = \frac{m\left(\pi_0 + \pi_1\right) - 2}{2(m-1)} \sigma_w^2 \neq \sigma_w^2.$$

  From (4.18), for $i = 0$, we can write

$$\text{E}\left(\bar{Y}_{0j}^* - \hat{\mu}_0^*\right)^2 = \left(1 - \frac{1}{k}\right)\left(\sigma_b^2 + \frac{\sigma_w^2}{m\pi_0}\right), \quad \text{using} \quad (4.13).$$

  Similarly, for $i = 1$, we have

$$\text{E}\left(\bar{Y}_{1j}^* - \hat{\mu}_1^*\right)^2 = \left(1 - \frac{1}{k}\right)\left(\sigma_b^2 + \frac{\sigma_w^2}{m\pi_1}\right).$$

Hence from (4.16), we can write

$$\begin{aligned}
\mathrm{E}\left(\mathrm{MSC}\right) &= \frac{m}{2(k-1)}\sum_{j=1}^{k}\left[\mathrm{E}\left(\bar{Y}_{0j}^{*}-\hat{\mu}_{0}^{*}\right)^{2}+\mathrm{E}\left(\bar{Y}_{1j}^{*}-\hat{\mu}_{1}^{*}\right)^{2}\right] \\
&= \frac{mk}{2(k-1)}\left(1-\frac{1}{k}\right)\left(2\sigma_{b}^{2}+\frac{\sigma_{w}^{2}}{m\pi_{0}}+\frac{\sigma_{w}^{2}}{m\pi_{1}}\right) \\
&= m\sigma_{b}^{2}+\left(\frac{\pi_{0}+\pi_{1}}{2\pi_{0}\pi_{1}}\right)\sigma_{w}^{2} \\
&= m\sigma_{b}^{2}+\frac{1}{\tilde{\pi}}\sigma_{w}^{2},
\end{aligned}$$

where $\tilde{\pi}$ is the harmonic mean of π_0 and π_1. Then

$$\begin{aligned}
\mathrm{E}\left(S_{b}^{2}\right) &= \frac{1}{m}\Big[\mathrm{E}\left(\mathrm{MSC}\right)-\mathrm{E}\left(\mathrm{MSW}\right)\Big] \\
&= \frac{1}{m}\left(m\sigma_{b}^{2}+\frac{1}{\tilde{\pi}}\sigma_{w}^{2}-\frac{m\left(\pi_{0}+\pi_{1}\right)-2}{2(m-1)}\sigma_{w}^{2}\right) \\
&= \sigma_{b}^{2}+\left(\frac{1}{\tilde{\pi}}-\frac{m\left(\pi_{0}+\pi_{1}\right)-2}{2(m-1)}\right)\frac{\sigma_{w}^{2}}{m}\neq\sigma_{b}^{2}.
\end{aligned}$$

Therefore, S_w^2 and S_b^2 are biased estimators for σ_w^2 and σ_b^2, respectively, with cluster mean imputation under MAR. Since, for $m > 0$ and $0 < \pi_0, \pi_1 < 1$, we have $0 < \pi_0 + \pi_1 < 2$ and $\tilde{\pi} < 1$; and

$$\frac{m\left(\pi_{0}+\pi_{1}\right)-2}{2(m-1)}<1 \qquad \text{and} \qquad \left(\frac{1}{\tilde{\pi}}-\frac{m\left(\pi_{0}+\pi_{1}\right)-2}{2(m-1)}\right)>0.$$

Hence, S_w^2 is downward biased for σ_w^2, whereas S_b^2 is upward biased for σ_b^2. Also since

$$\frac{m(\pi_{0}+\pi_{1})-2}{2(m-1)}=\frac{(\pi_{0}+\pi_{1})-\frac{2}{m}}{2-\frac{2}{m}}\rightarrow\bar{\pi}\quad\text{as}\quad m\rightarrow\infty$$

and

$$\frac{1}{m}\left(\frac{1}{\tilde{\pi}} - \frac{m\left(\pi_0+\pi_1\right)-2}{2(m-1)}\right) = \left(\frac{1}{m\tilde{\pi}} - \frac{\left(\pi_0+\pi_1\right)-2}{2(m-1)}\right) \to 0 \quad \text{as} \quad m \to \infty,$$

we deduce $\mathrm{E}\left(S_w^2\right) \approx \bar{\pi}\sigma_w^2$ and $\mathrm{E}\left(S_b^2\right) \approx \sigma_b^2$ for large m. Therefore, with large cluster size, σ_w^2 is underestimated by a factor approximately equal to the average of the follow-up rates π_0 and π_1, and S_b^2 is approximately unbiased for σ_b^2.

4.4 Simulation study I

A simulation study was conducted to investigate the consequence of cluster mean imputation to the unbiasedness of the ANOVA estimators of variance components. The simulation study was designed to mimic data from a worksite obesity intervention trial used by Taljaard *et al.* [17]. In this simulation study, we considered balanced CRTs.

4.4.1 Data generation

Data was generated using the model defined in equation (3.1). Parameters were fixed as $\sigma^2 = 225$, $\mu_1 = \mu_2 = 75$, and $\rho = 0.1$. Parameters that were varied in generating the data include the number of clusters in each group $k = (5, 10, 15, 20, 30)$ and the cluster size $m = (30, 50, 100, 250)$. The values of σ_w^2 and σ_b^2 were then determined by the relation $\rho = \sigma_b^2/(\sigma_b^2 + \sigma_w^2)$. Missing data indicators R_{ijl} were generated under the two different examples of MCAR (MCAR1 and MCAR2, defined in equations (4.4) and (4.5), respectively) and under MAR, defined in equation (4.7). For MCAR1, we

set constant follow-up rate $\pi = 0.7$. For MCAR2, we set variable but independent follow-up rates as $\pi_{ij} \sim \text{Uniform}(0.4, 1)$. For MAR, we fixed follow-up rates $\pi_0 = 0.6$ in the control group and $\pi_1 = 0.8$ in the intervention group. We chose follow-up rates under MCAR2 and MAR so that the average follow-up rates in both situations were equal to the constant follow-up rate under MCAR1. Missing values indicators were then imposed on each generated full data to get the incomplete data.

4.4.2 Imputation and analysis

We imputed missing outcomes with cluster mean imputation. The resulting imputed datasets were then used to evaluate S_w^2 and S_b^2, the ANOVA estimators of σ_w^2 and σ_b^2, respectively. Note that, the ANOVA estimators S_w^2 and S_b^2 with full data are unbiased for σ_w^2 and σ_b^2, respectively.

4.4.3 Results

The average estimates of the variance components over 1000 simulation runs are presented in Table 4.1. The average estimates of the within-cluster variance were much lower compared to the true value under MCAR1, MCAR2 and MAR, as expected. As we showed analytically, a better estimates of the within-cluster variance can be obtained by dividing the within-cluster variance estimates in Table 4.1 by the average follow-up rates. The between-cluster variance was overestimated for small cluster size. However, the between-cluster variance estimates tended to be close to the true value as the cluster size went up. These results support our derived analytical results in Section 4.3, where

Table 4.1: Average estimates of within-cluster variance (σ_w^2) and between-cluster variance (σ_b^2) over 1000 simulation runs using cluster mean imputation for missing outcomes under (a) MCAR1 with $\pi = 0.7$ (b) MCAR2 with $\pi_{ij} \sim \text{Uniform}(0.4, 1)$, and (c) MAR with $\pi_0 = 0.6$ and $\pi_1 = 0.8$. The true values are $\sigma_w^2 = 202.5$ and $\sigma_b^2 = 22.5$.

	Within-cluster variance ($\sigma_w^2 = 202.5$)				Between-cluster variance ($\sigma_b^2 = 22.5$)			
	$m = 30$	50	100	250	$m = 30$	50	100	250
(a) MCAR1								
$k = 5$	139.92	140.59	141.34	141.78	27.32	25.22	24.44	23.32
10	139.77	140.99	141.15	141.60	27.78	25.89	23.88	22.80
15	139.89	140.53	141.15	141.62	27.65	25.71	24.37	23.13
20	139.75	140.69	141.41	141.56	27.97	25.93	24.21	23.14
30	139.65	140.62	141.17	141.52	27.97	25.49	24.22	23.28
(b) MCAR2								
$k = 5$	140.20	141.14	141.58	141.49	28.41	26.45	24.15	23.08
10	139.93	140.73	141.42	141.32	28.25	26.24	23.97	23.36
15	139.71	140.46	140.88	141.69	28.47	26.51	23.92	23.19
20	139.65	140.93	140.97	141.56	28.57	26.14	24.44	23.32
30	139.58	140.62	141.23	141.62	28.32	26.05	24.35	23.10
(c) MAR								
$k = 5$	139.65	140.67	141.37	141.73	27.58	25.42	24.53	23.31
10	139.94	140.84	141.23	141.53	27.89	25.84	23.92	22.77
15	140.05	140.70	141.13	141.57	27.87	25.82	24.43	23.15
20	139.59	140.62	141.38	141.55	28.27	26.07	24.34	23.19
30	139.59	140.71	141.15	141.52	28.17	25.61	24.26	23.32

we showed that the ANOVA estimators of the variance components are biased under particular MCAR and MAR mechanisms with cluster mean imputation. For fixed m and follow-up rates, the average variance components estimates across simulations remained almost same as the value of k increases since the expected value of S_w^2 and S_b^2 do not depend on k.

4.5 Cluster-level $t-$test, adjusted $t-$test and LMM under balanced CRT

We described the cluster-level $t-$test and adjusted $t-$test in general in Section 3.2 and Section 3.3, respectively. In the case of balanced CRT with $k_i = k$ and $m_{ij} = m$, the cluster-level $t-$test is given by

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_0}{S\sqrt{2/k}} \sim t_{(2k-2)}. \tag{4.19}$$

Referring to the Section 3.3, it can be shown that, for balanced CRT, $\tilde{\mu}_i = \hat{\mu}_i$, $m_0 = A_0 = A_1 = m$, and $\widehat{\text{VIF}}_0 = \widehat{\text{VIF}}_1 = 1 + (m-1)\hat{\rho}$. Hence, the adjusted $t-$test is given by

$$t_A = \frac{\hat{\mu}_1 - \hat{\mu}_0}{\widehat{\text{SE}}\,(\hat{\mu}_1 - \hat{\mu}_0)} \sim t_{2k-2} \tag{4.20}$$

where

$$\begin{aligned}
\widehat{\text{SE}}\,(\hat{\mu}_1 - \hat{\mu}_0) &= \sqrt{\frac{2S_P^2}{km}[1 + (m-1)\hat{\rho}]} \\
&= \sqrt{\frac{2}{km}\left(S_P^2 + (m-1)\hat{\rho}S_P^2\right)} \\
&= \sqrt{\frac{2}{km}\left(S_w^2 + S_b^2 + (m-1)S_b^2\right)} \\
&= \sqrt{\frac{2}{km}\left(S_w^2 + mS_b^2\right)} \\
&= \sqrt{\frac{2}{km}\text{MSC}} \\
&= S\sqrt{2/k},
\end{aligned}$$

which is exactly the same standard error of $\hat{\mu}_1 - \hat{\mu}_0$ used in test (4.19). This proves that the adjusted $t-$test and the cluster-level $t-$test are identical for balanced CRTs. In the case of missing outcomes in a balanced CRT, these two tests are identical with

cluster mean imputation since after imputation the cluster sizes becomes constant and the cluster means remain unchanged. Also the cluster-level $t-$test with CRA and with cluster mean imputation are identical because cluster mean imputation does not change the cluster means. Therefore, with a balanced design, the cluster-level $t-$test with CRA is identical with cluster-level $t-$test and the adjusted $t-$test with cluster mean imputation.

As we described in Section 1.3.2, LMM is used as an individual-level analysis for continuous outcomes. In LMM, the estimated mean of the ith intervention group, denoted by $\hat{\mu}_i^{\text{lmm}}$, is calculated as

$$\hat{\mu}_i^{\text{lmm}} = \frac{\sum_{j=1}^{k_i} v_{ij}\bar{Y}_{ij}}{\sum_{j=1}^{k_i} v_{ij}}, \qquad \text{where} \qquad v_{ij} = \left(S_b^2 + S_w^2/m_{ij}\right)^{-1}$$

In the case of balanced CRT, this mean is exactly the same as the intervention groups means calculated in cluster-level $t-$test and adjusted $t-$test.

4.6 Simulation study II

A simulation study was conducted to investigate the impact of CRA and cluster mean imputation for missing outcomes on the validity and power of cluster-level $t-$test, adjusted $t-$test and LMM under MCAR and MAR. We also investigate the power of these approaches when cluster follow-up rates are highly variable. In this simulation study, we considered balanced CRTs.

4.6.1 Data generation

Data was generated using the model defined in equation (3.1). Parameters were fixed as $\sigma^2 = 225$, $\mu_1 = \mu_2 = 75$ to report Type I error and $\mu_2 - \mu_1 = 5$ to report power values. Parameters that were varied in generating the data include the number of clusters in each group $k = (5, 10, 15, 20, 30)$, the cluster size $m = (30, 50, 100, 250)$, and the intraclass correlation coefficient $\rho = (0.001, 0.01, 0.05, 0.1)$. The values of σ_w^2 and σ_b^2 were then determined by the relation $\rho = \sigma_b^2/(\sigma_b^2 + \sigma_w^2)$. Missing data indicators R_{ijl} were generated under MCAR1 (defined in equation (4.4)) with constant follow-up rate $\pi = 0.7$, under MCAR2 (defined in equation (4.5)) with variable but independent follow-up rates $\pi_{ij} \sim \text{Uniform}(0.4, 1)$, and under MAR (defined in equation (4.7)) with follow-up rates $\pi_0 = 0.6$ in the control group and $\pi_1 = 0.8$ in the intervention group. We chose follow-up rates under MCAR2 and MAR so that the average follow-up rates on both situations were equal to the constant follow-up rate under MCAR1. To investigate the power with higher variation among cluster sizes, we considered another example of MCAR, here referred to as MCAR3, where cluster follow-up rates $\pi_{ij} \sim \text{Uniform}(0, 1)$ and each cluster has at least one observed value.

4.6.2 Imputation and analysis

At first, each full dataset was used to test the hypothesis of no intervention, with $\alpha = 0.05$, for comparison purpose using both cluster-level analysis and individual-level analysis. We used cluster-level $t-$test for cluster-level analysis, whereas $z-$test and Wald $t-$test (with Satterthwaite's approximation for DF) were used for individual-level

analysis with LMM. After imposing missing values, the same analyses were performed using CRA. We explored adjusted $t-$test only for CRA since after imputation the cluster sizes become constant in which case this test is identical with the cluster-level $t-$test. Then we imputed missing outcomes with observed cluster mean. The resulted imputed datasets were then analysed using LMM to test the same hypothesis.

4.6.3 Results: Type I error

The cluster-level $t-$test and the adjusted $t-$test gave Type I error rate very close to nominal level (0.05) using full data, CRA and cluster mean imputation (results not presented here). This is expected because cluster-level $t-$test is robust in terms of Type I error rate even with small number of clusters in each group [5] and, in the case of balanced CRT, the cluster-level $t-$test and the adjusted $t-$test are equivalent. Also Taljaard *et al.* [17] showed that adjusted $t-$test gives acceptable Type I error rate using CRA and cluster mean imputation under MCAR. Because of this, the results of Type I error rate of cluster-level $t-$test and adjusted $t-$test are not presented in this thesis. Empirical Type I error rates of $z-$test and Wald $t-$test over 1000 simulation runs using full data, CRA and cluster mean imputation under MCAR1 are presented in Table 4.2. The $z-$test tended to have inflated Type I error rates for small number of clusters ($k \leq 10$) using full data, CRA and cluster mean imputation for missing values. This is due to the fact that inferences for fixed effects using $z-$test are based on their asymptotic distribution which is insufficient for smaller number of clusters in each group. The error rates were close to the nominal level for higher number of clusters ($k \geq 15$) in each group. We used Satterthwaite's approximation to calculate the degrees of freedom of Wald $t-$test for intervention effect since it is the only fixed effect in our

analysis. The Wald $t-$test gave Type I error rates close to the nominal level at all considered values of k, m and ρ. Both tests resulted in acceptable Type I error rate for $k > 15$ and so the results are not presented in the table for $k > 15$. Qualitatively similar results of Type I error were observed under MCAR2 and MAR (see Table A1 and Table A2, respectively, in **Appendix A**.)

4.6.4 Results: power values

Empirical power estimates of the cluster-level $t-$test , adjusted $t-$test and Wald $t-$test for intervention effect are shown in Table 4.3 using full data, CRA and cluster mean imputation for missing outcomes under MCAR1. The power values of adjusted $t-$test using full data are not presented because the cluster-level $t-$test and the adjusted $t-$test are identical under balanced CRTs. Also, since the cluster-level $t-$test with CRA is identical with cluster-level $t-$test and the adjusted $t-$test with cluster mean imputation, the results of these tests are not presented under cluster mean imputation. The power values for $z-$test were not calculated since it gives inflated Type I error rates for small number of clusters in each group (see Table 4.2).

As expected, the power values using CRA went down compared to that of using full data. All three considered tests tended to have similar power with CRA. It was not surprising due to the fact that the variation among cluster sizes was very low with a constant follow-up rate. The LMM with cluster mean imputation did not gain extra power compared to that of LMM using CRA. The power values showed an increasing trend at all values of ρ for higher values of k and m. However, they increased more rapidly for higher values of k compared to that of higher values of m. It was expected

as the power depends on the number of clusters in each group to a great extent than on the cluster size [2]. Qualitatively similar results of power values are observed under MCAR2 and under MAR (see Table A3 and Table A4, respectively). One of the reasons for this might be the selection of follow-up rates in our simulation setup, where we chose the follow-up rates in such a way that the average follow-up rates remained same under MCAR1, MCAR2 and MAR. A large variation in cluster sizes could result in improved power for adjusted $t-$test and LMM.

The empirical power values of cluster-level $t-$test, adjusted $t-$test and LMM using CRA under MCAR3 are presented in Table 4.4. The adjusted $t-$test and LMM tended to have higher power compared to the power of cluster-level $t-$test as the value of ρ went down. This is due to the fact that cluster sizes are taken as weights in adjusted $t-$test, and in LMM weights are a function of cluster size as well as variance components. On the other hand, in cluster-level $t-$test, equal weights are given to the cluster means ignoring the variation in cluster sizes. Therefore, in cluster-level $t-$test with a small value of ρ, which measures the similarity between two observations in the same cluster, cluster mean with very few observation may not accurately represent the true mean of that cluster.

Table 4.2: Empirical Type I error rate over 1000 simulation runs of LMM with the $z-$test and the Wald $t-$test (using Satterthwaite's approximation for degrees freedom) for intervention effect using full data, CRA and cluster mean imputation under MCAR1.

k	m	ρ	Full data		CRA		Cluster mean imputation	
			$z-$ test	Wald $t-$test	$z-$ test	Wald $t-$test	$z-$ test	Wald $t-$test
5	30	0.01	4.8	3.6	6.4	4.7	**9.3**	6.2
		0.05	**7.6**	4.2	**8.3**	5.0	**9.1**	5.5
		0.10	**8.3**	5.1	**8.5**	4.7	**8.7**	4.8
	50	0.01	6.8	4.6	5.9	4.0	**7.5**	4.4
		0.05	**9.0**	5.8	**7.7**	5.1	**7.5**	5.0
		0.10	**8.9**	5.2	**8.2**	4.7	**7.9**	4.6
	100	0.01	6.2	4.3	6.7	4.8	**7.9**	4.9
		0.05	**7.9**	3.7	**8.4**	4.8	**8.7**	5.0
		0.10	**7.9**	4.1	**8.8**	4.7	**8.7**	4.8
	250	0.01	**8.5**	5.4	6.9	4.3	**8.7**	5.1
		0.05	**8.6**	4.6	**9.3**	5.0	**9.4**	5.1
		0.10	**8.6**	4.4	**9.3**	5.0	**9.2**	5.0
10	30	0.01	5.8	5.0	4.4	3.3	5.8	4.5
		0.05	**7.6**	6.4	6.6	5.0	6.8	5.3
		0.10	**7.4**	6.0	**7.1**	4.9	**7.5**	4.8
	50	0.01	5.5	4.8	6.3	5.7	6.8	5.1
		0.05	**7.4**	5.3	6.6	5.3	6.7	5.6
		0.10	**7.3**	5.6	6.4	4.7	6.4	5.0
	100	0.01	5.0	3.7	6.8	5.1	**8.3**	5.9
		0.05	5.9	4.0	**7.7**	5.8	**7.4**	5.9
		0.10	5.9	4.6	6.9	5.5	6.8	5.5
	250	0.01	4.7	3.4	**7.0**	5.1	**9.1**	**7.4**
		0.05	5.6	4.4	**8.1**	5.8	**8.1**	5.9
		0.10	5.7	4.6	**7.7**	5.5	**7.6**	5.5
15	30	0.01	5.5	4.9	5.1	4.4	5.6	4.3
		0.05	6.1	5.5	5.1	4.8	5.4	4.8
		0.10	5.9	5.3	5.6	4.9	6.0	4.9
	50	0.01	6.4	5.4	6.7	6.1	5.3	4.8
		0.05	5.5	4.9	6.0	5.2	6.2	5.5
		0.10	5.7	4.6	6.2	5.3	6.1	5.5
	100	0.01	5.1	4.1	6.1	5.4	6.3	5.4
		0.05	5.4	4.4	5.6	4.6	5.6	4.5
		0.10	5.6	4.4	4.7	4.1	4.7	4.0

Table 4.3: Empirical power values of the cluster-level $t-$test, adjusted $t-$test and LMM with Wald $t-$test for intervention effect over 1000 simulation runs using full data, CRA and cluster mean imputation for missing outcomes under MCAR1.

k	m	ρ	Full data		CRA			LMM
			Cluster level $t-$test	LMM approach	Cluster level $t-$test	Adjusted $t-$test	LMM approach	with cluster mean imputation
5	30	0.01	63.8	60.9	49.4	49.4	48.9	49.4
		0.05	38.0	38.2	33.4	33.9	33.6	33.4
		0.10	25.6	26.6	23.5	23.6	23.6	23.5
	50	0.01	76.6	74.8	66.7	67.1	67.6	66.6
		0.05	42.9	39.6	39.3	39.8	39.7	39.3
		0.10	28.3	25.0	27.4	28.0	27.4	27.4
	100	0.01	90.5	89.2	83.2	83.4	85.7	83.2
		0.05	48.7	46.8	43.3	43.2	43.3	43.3
		0.10	28.7	29.3	27.7	27.6	27.9	27.7
	250	0.01	97.3	97.2	95.5	95.7	95.0	95.5
		0.05	54.2	51.7	51.9	51.8	51.8	51.9
		0.10	31.6	33.0	31.4	31.6	31.4	31.4
10	30	0.01	92.0	93.4	83.5	83.9	83.0	83.4
		0.05	68.2	71.0	62.9	62.6	62.9	62.9
		0.10	50.7	51.2	47.2	46.8	46.9	47.2
	50	0.01	98.4	97.7	95.6	95.4	94.6	95.6
		0.05	76.4	77.6	70.7	70.7	70.6	70.7
		0.10	52.1	54.8	49.5	48.9	49.7	49.5
	100	0.01	99.7	99.8	99.4	99.4	99.5	99.4
		0.05	82.7	84.1	80.1	80.1	80.3	80.1
		0.10	57.3	58.9	55.0	55.0	54.9	55.0
20	30	0.01	99.8	100	98.6	98.5	99.1	98.6
		0.05	94.3	96.0	90.4	90.4	90.7	90.4
		0.10	80.0	80.9	75.8	74.9	75.4	75.8
	50	0.01	100	100	100	100	100	100
		0.05	97.0	97.1	96.3	96.0	96.2	96.3
		0.10	84.0	86.1	83.0	82.8	82.9	83.0
	100	0.01	100	100	100	100	100	100
		0.05	98.4	98.9	97.8	98.1	97.8	97.8
		0.10	87.3	87.0	86.1	86.1	86.1	86.1

Table 4.4: Empirical power of the cluster level $t-$test, adjusted $t-$test and Wald $t-$test using CRA under MCAR3.

k	m	ρ	Cluster-level $t-$test	Adjusted $t-$test	LMM approach
5	50	0.001	40.4	50.7	54.6
		0.005	38.8	49.4	52.4
		0.010	37.1	46.2	48.3
		0.050	25.2	28.6	29.6
		0.100	18.1	19.7	20.9
	100	0.001	62.2	80.3	81.1
		0.005	58.3	74.5	75.1
		0.010	55.1	67.4	68.1
		0.050	36.3	39.3	40.9
		0.100	25.8	25.5	27.3
	250	0.001	85.3	98.7	98.3
		0.005	82.9	94.5	94.9
		0.010	76.9	87.0	88.8
		0.050	43.2	42.1	45.5
		0.100	28.0	25.2	28.4
10	30	0.001	48.4	71.1	73.4
		0.005	46.9	68.3	69.9
		0.010	45.9	66.3	66.1
		0.050	38.0	48.1	48.5
		0.100	31.3	35.6	35.9
	100	0.001	83.9	99.4	99.6
		0.005	81.8	98.7	98.6
		0.010	80.1	96.3	96.6
		0.020	76.1	89.7	90.5
		0.050	61.2	67.6	70.8
		0.100	45.1	45.7	50.0
15	30	0.001	65.3	88.3	89.6
		0.005	64.4	87.5	87.7
		0.010	63.1	85.9	86.5
		0.020	60.7	81.5	82.5
		0.050	55.7	69.3	71.5
		0.100	48.1	53.7	56.8

4.7 Summary

First, this chapter showed analytically and through simulations that the ANOVA estimators of the variance components are biased with cluster mean imputation. In the case of large cluster size, within-cluster variance is underestimated by a factor approximately equal to the average of the follow-up rates, and the between -cluster variance estimator is approximately unbiased. Hence the estimate of total variance, which is the sum of between-cluster variance and within-cluster variance, is also biased, and the estimate of ICC is also biased since it is calculated as the proportion of total variability that is attributed to between-cluster variability. Therefore, we do not recommend cluster mean imputation, since the variance components and ICC are often of interest.

Second, this chapter showed analytically that cluster-level $t-$test and adjusted $t-$test are identical under balanced CRTs. In the case of missing outcome in a balanced CRT, cluster-level $t-$test with CRA is identical with cluster-level $t-$test and adjusted $t-$test with cluster mean imputation.

Third, in LMM with small number of clusters in each intervention group, the Wald $t-$test with Satterthwaite's approximation for DF yielded acceptable type I error rate, whereas the $z-$test tended to have inflated type I error. However, both tests gave acceptable type I error with large number of clusters in each intervention groups. When cluster sizes do not vary largely, cluster-level $t-$test, adjusted $t-$test, and LMM with Wald $t-$test gave similar power with full data, CRA and cluster mean imputation under the considered examples of MCAR and MAR. Therefore, in this situation, cluster-level $t-$test could be an attractive option because of its simplicity compared to the other

two test procedures. However, if the cluster sizes vary considerably, which partly could arise because of varying follow-up rates between clusters, adjusted $t-$test and LMM gave better power compared to that of cluster-level $t-$test using CRA at small values of ICC.

Chapter 5

Research Paper I

Title: Missing continuous outcomes under covariate dependent missingness in cluster randomised trials.

Author(s): Anower Hossain, Karla DiazOrdaz and Jonathan W. Bartlett.

Journal Name: Statistical Methods in Medical Research.

Type of publication: Research paper.

Stage of Publication: Volume 26, Issue 3, June 2017, Pages 1543-1562.

DOI: 10.1177/0962280216648357

URL: `http://smm.sagepub.com/content/early/2016/05/12/0962280216648357.full.pdf`

Academic peer-reviewed: Yes.

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
www.lshtm.ac.uk

Registry
T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Anower Hossain (AH)
Principal Supervisor	Karla DiazOrdaz
Thesis Title	Missing data in cluster randomised trials

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?	Statistical Methods in Medical Research		
When was the work published?	14 May 2016		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	NA		
Have you retained the copyright for the work?*	Yes	Was the work subject to academic peer review?	Yes

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	AH identified the research questions, designed and run the simulation study, and interpreted the results with supervision from the supervisors. AH wrote the initial draft of the manuscript and then revised it based on the feedback received from the supervisors. All authors read and approved the final manuscript.

Student Signature: Hossain **Date:** 29.08.17

Supervisor Signature: **Date:** 29/08/17

Summary of Research Paper I

Title: Missing continuous outcomes under covariate dependent missingness in cluster randomised trials.

This research paper investigates the validity of unadjusted and adjusted cluster-level analyses and LMM for analysing CRTs when the outcomes are continuous and only outcomes are missing under CDM assumption. The methods of CRA and MI are used to handle missing outcomes. We show that the unadjusted and adjusted cluster-level analyses are in general biased unless the intervention groups have the same missingness mechanism and the same covariate effects in the data generating model for the outcome, which is arguably unlikely to hold in practice. The LMM using CRA adjusted for covariates such that the CDM assumption holds give unbiased estimates of intervention effect regardless of whether the intervention groups had the same missingness mechanism, and whether there is an interaction between intervention and baseline covariate in the data generating model for the outcome, provided that such interaction is included in the model when required.

We compare the results of LMM using CRA adjusted for covariates such that the CDM assumption holds with the results of MI. We find that there is no gain in terms of bias or efficiency of the estimates using MI over CRA adjusted for covariates, when both approaches used the same functional form of the same set of baseline covariates and the same modelling assumptions. In conclusion, in the absence of auxiliary variables,

LMM using CRA adjusted for covariates such that the CDM assumption holds can be recommended as the primary analysis approach for CRTs with missing outcome if one is willing to make the CDM assumption for outcomes.

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2017, Vol. 26(3) 1543–1562

DOI: 10.1177/0962280216648357
journals.sagepub.com/home/smm
SAGE

Missing continuous outcomes under covariate dependent missingness in cluster randomised trials

Anower Hossain,[1] Karla Diaz-Ordaz[1] and Jonathan W Bartlett[2]

Abstract

Attrition is a common occurrence in cluster randomised trials which leads to missing outcome data. Two approaches for analysing such trials are cluster-level analysis and individual-level analysis. This paper compares the performance of unadjusted cluster-level analysis, baseline covariate adjusted cluster-level analysis and linear mixed model analysis, under baseline covariate dependent missingness in continuous outcomes, in terms of bias, average estimated standard error and coverage probability. The methods of complete records analysis and multiple imputation are used to handle the missing outcome data. We considered four scenarios, with the missingness mechanism and baseline covariate effect on outcome either the same or different between intervention groups. We show that both unadjusted cluster-level analysis and baseline covariate adjusted cluster-level analysis give unbiased estimates of the intervention effect only if both intervention groups have the same missingness mechanisms and there is no interaction between baseline covariate and intervention group. Linear mixed model and multiple imputation give unbiased estimates under all four considered scenarios, provided that an interaction of intervention and baseline covariate is included in the model when appropriate. Cluster mean imputation has been proposed as a valid approach for handling missing outcomes in cluster randomised trials. We show that cluster mean imputation only gives unbiased estimates when missingness mechanism is the same between the intervention groups and there is no interaction between baseline covariate and intervention group. Multiple imputation shows overcoverage for small number of clusters in each intervention group.

Keywords

Cluster randomised trials, missing outcome data, covariate dependent missingness, multiple imputation, complete records analysis

[1]Department of Medical Statistics, London School of Hygiene & Tropical Medicine (LSHTM), London, UK
[2]Statistical Innovation Group, AstraZeneca, Cambridge, UK

Corresponding author:
Anower Hossain, Department of Medical Statistics, London School of Hygiene & Tropical Medicine (LSHTM), Keppel Street, London, WC1E 7HT, UK.
Email: anower.hossain@lshtm.ac.uk

1 Introduction

In cluster randomised trials (CRTs), identifiable clusters of individuals such as villages, schools, medical practices – rather than individuals – are randomly allocated to each of intervention and control groups, while individual-level outcomes of interest are observed within each cluster. The number of clusters and/or the cluster sizes in each intervention group might be different. CRTs with equal number of clusters in each intervention group with constant cluster size are known as balanced CRTs. One important characteristic of CRTs is that the outcomes of individuals within the same cluster may exhibit more similarity compared to the outcomes of individuals in the other clusters, which is quantified by the intraclass correlation coefficient (ICC), denoted by ρ. In practice, the value of ICC typically ranges from 0.001 to 0.05 and it is rare for clinical outcomes to have ICC above 0.1.[1] Small values of ICC can lead to substantial variance inflation factors and should not be ignored.[2,3] CRTs are being increasingly used in the fields of health promotion and health service research. Reasons for such popularity include the nature of intervention that itself may dictate its application at the cluster level, less risk of intervention contamination and administrative convenience.[4] It is well known that the power and precision of CRTs are lower relative to trials that individually randomise the same number of individuals.[2] In spite of this, the advantages associated with CRTs are perceived by researchers to outweigh the potential loss of statistical power and precision in some situations.

Attrition is a common problem for CRTs, leading to missing outcome data. This not only reduces the statistical power of the study but may result in biased intervention effect estimates.[5] Handling missing data in CRTs is complicated by the fact that data are clustered. Inadequate handling of the missing data may result in misleading inferences.[6] A systematic review[7] revealed that, among all CRTs published in English in 2011, 72% of trials had missing values either in outcomes or in covariates or in both. Among them only 34% of CRTs reported how they handled missing data. One of the reasons may be that the methodological development for dealing with missing data in CRTs has been relatively slow in spite of the increasing popularity of CRTs. Cluster mean imputation has been suggested as a valid approach for handling missing outcome data in CRTs.[8]

The impact of missing data on estimation and inference of a parameter of interest depends on the missing data mechanism, the method used to handle the missing data, and the choice of statistical methods used for data analysis. In this paper, we study the validity of three analysis methods – unadjusted cluster-level analysis, adjusted cluster-level analysis and linear mixed model (LMM) – when there is missingness in the continuous outcome, and this missingness depends on baseline covariates, and conditional on these baseline covariates, not on the outcomes itself. We compare the performance of these methods on complete records and multiply imputed datasets. In addition, we investigate the validity of cluster mean imputation, as proposed by Taljaard et al.,[8] under the same missingness assumption.

This paper is organised as follows. Section 2 presents a brief review of the approaches to the analysis of CRTs with complete data. In Section 3, the assumed missingness mechanism for CRTs is described. Section 4 describes methods of handling missing data in CRTs. In Section 5, we investigate the validity of complete records analysis of CRTs. Section 6 describes a simulation study and presents the results. We conclude the study with some discussion in Section 7.

2 Analysis of CRTs with complete data

We begin by describing the two broad approaches to the analysis of CRTs in the absence of missing data. These are cluster-level analysis and individual-level analysis.

2.1 Cluster-level analysis

Cluster-level analysis can be done in two ways: unadjusted cluster-level analysis and baseline covariate adjusted cluster-level analysis. This approach can be explained as a two-stage process. In the first stage of unadjusted analysis, a relevant summary measure of outcomes is calculated for each cluster. Then, in the second stage, the cluster-specific summary measures of the control and intervention groups obtained in the first stage are compared using appropriate statistical methods. The most common one is the standard *t*-test for two independent samples (here referred to as cluster-level *t*-test) with degrees of freedom (DF) equal to the total number of clusters in the study minus two. The basis of using this test is that the resulting summary measures are statistically independent, which is a consequence of the clusters being independent of each other. In the case of baseline covariate adjusted analysis, an individual-level regression analysis is carried out at the first stage including all covariates as explanatory variables, except for the intervention indicator, and ignoring the clustering of the data.[4,9] The individual level residuals from the first-stage model are then used to calculate the cluster-specific summary measures for the control group and the intervention group, which are then compared using cluster-level *t*-test in the second stage of analysis to evaluate the intervention effect adjusted for baseline covariates. The main purposes of adjusting for baseline covariates are to increase the credibility of the trial findings by demonstrating that any observed intervention effect is not attributed to the possible imbalance between the intervention groups in term of baseline covariates and to improve the statistical power.[10]

2.2 Individual-level analysis

In individual-level analysis, a regression model is fitted to the individual-level outcomes, allowing for the fact that observations within the same cluster are correlated. LMM is widely used as individual-level analysis for CRTs with continuous outcomes. The LMM takes into account between-cluster variability using cluster-level effects which are assumed to follow a specified probability distribution. The parameters of that distribution are estimated using maximum likelihood methods together with intervention effect and other covariates effects. Generalised estimating equations are an alternative approach, but for continuous outcomes and an exchangeable correlation matrix, estimates are identical to those from LMM with a random intercept.[11]

The adjusted *t*-test, proposed by Donner and Klar,[2] is an alternative approach to test the intervention effect for quantitative outcomes, which involves calculating the mean of the individual outcome values in each intervention group. These means are then compared using a *t*-test in which the standard error (SE) is adjusted to account for the intracluster correlation. The adjusted *t*-test and the cluster-level *t*-test are identical for balanced CRTs.

3 Missingness mechanism assumptions for CRTs

In this paper, we will consider the common setting where the outcomes are continuous, and only outcomes are missing. In statistical analysis, if there are missing values, an assumption must be made about the missingness mechanism, which refers to the relationship between missingness and the underlying values of the variables in the data.[12] According to Rubin's framework,[13] a missingness mechanism can be classified as (i) missing completely at random (MCAR), where the probability of a value being missing is independent of the observed and unobserved data; (ii) missing at random (MAR), where conditioning on the observed data, the probability of a value being missing is independent of the unobserved data; and (iii) missing not at random (MNAR), where the probability of value being missing depends on both observed and unobserved data.

In CRTs, an assumption that may sometimes be plausible is that missingness in outcomes depends on covariates measured at baseline and conditional on these baseline covariates, not on the outcome itself. We refer to this as covariate dependent missingness (CDM). For example, blood pressure outcome data could be CDM if missingness in blood pressure measurement depends on covariates (e.g. age, BMI or weight), but given these, not on the blood pressure measurement itself. CDM is an example of a MAR mechanism when covariates are fully observed.

Let Y_{ijl} be a continuous outcome of interest for the lth ($l = 1, 2, \ldots, m_{ij}$) individual in the jth ($j = 1, 2, \ldots, k_i$) cluster of the intervention group i ($i = 1, 2$), where $i = 1$ corresponds to control group and $i = 2$ corresponds to intervention group. We assume that the Y_{ijl} follow a LMM given by

$$Y_{ijl} = \alpha_i + \beta_i X_{ijl} + \delta_{ij} + \epsilon_{ijl} \tag{1}$$

where α_i is a constant for ith intervention group, X_{ijl} is a baseline covariate value for (ijl)th individual, β_i is the effect of baseline covariate X on Y in intervention group i, δ_{ij} is the (ij)th cluster effect and ϵ_{ijl} is the individual error term. We also assume that the cluster effect (δ_{ij}) and the individual error (ϵ_{ijl}) are statistically independent, and $\mathrm{E}(\delta_{ij}) = 0$, $\mathrm{Var}(\delta_{ij}) = \sigma_b^2$ and $\mathrm{E}(\epsilon_{ijl}) = 0$, $\mathrm{Var}(\epsilon_{ijl}) = \sigma_w^2$, where σ_b^2 and σ_w^2 are the between-cluster variance and within-cluster variance, respectively. Later we will sometimes make normality assumptions on these random effects/random errors. Suppose the baseline covariate X has mean μ_x. Then

$$\mathrm{E}(\bar{Y}_i) = \alpha_i + \beta_i \mu_x = \mu_i$$

where $\bar{Y}_i = (1/k_i) \sum_{j=1}^{k_i} (1/m_{ij}) \sum_{l=1}^{m_{ij}} Y_{ijl} = (1/k_i) \sum_{j=1}^{k_i} \bar{Y}_{ij}$. Here, $\bar{Y}_i$ and $\bar{Y}_{ij}$ are the mean outcome of the ith intervention group and the (ij)th cluster, respectively. With complete data, the cluster-level analysis estimate of the intervention effect, say $\hat{\theta}$, is then calculated as

$$\hat{\theta} = \bar{Y}_1 - \bar{Y}_2$$

With complete data, this estimator is unbiased for the true intervention effect, that is

$$\mathrm{E}(\hat{\theta}) = \mu_1 - \mu_2$$

Suppose there are some missing values for outcome Y. Define a missing data indicator R_{ijl} such that

$$R_{ijl} = \begin{cases} 1, & \text{if } Y_{ijl} \text{ is observed} \\ 0, & \text{if } Y_{ijl} \text{ is missing} \end{cases}$$

Then $\sum_{l=1}^{m_{ij}} R_{ijl}$ is the number of observed outcomes in the (ij)th cluster. The CDM assumption can then be expressed as

$$P(R_{ijl} = 0 | \mathbf{Y}_{ij}, \mathbf{X}_{ij}) = P(R_{ijl} = 0 | X_{ijl})$$

where $\mathbf{Y}_{ij} = (Y_{ij1}, Y_{ij2}, \ldots, Y_{ijm_{ij}})$ and $\mathbf{X}_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijm_{ij}})$ are the vectors of the outcomes and the baseline covariate values, respectively, in the (ij)th cluster. In other words, the missingness of the (ijl)th individual's outcome Y_{ijl} depends only on that individual's baseline covariate value X_{ijl}.

4 Methods of handling missing data in CRTs

Common approaches for handling missing data in CRTs include complete records analysis (CRA), single imputation and multiple imputation (MI). This section describes these approaches. In this paper, we focused on CRA and MI since they are the most commonly used methods for handling missing data.

4.1 CRA

In CRA, often referred to as complete case analysis, only individuals with outcome observed are considered in the analysis, while individuals with missing outcome are excluded. It is widely used because of its simplicity and is usually the default method of most statistical packages. It is well known that CRA is valid if data are MCAR or if missingness is independent of the outcome, conditional on covariates.[12] Likelihood-based CRA is valid under MAR, if missingness is only in the outcome and all predictors of missingness are conditioned on in the model.[12] CRA is also valid under MNAR mechanisms where missingness in a covariate is dependent on the value of that covariate, but is conditionally independent of outcome.[14,15]

4.2 Single imputation

Single imputation imputes a single value for each missing outcome and creates a complete dataset. In general, single imputation is not recommended, since estimates of uncertainty are biased downwards, leading to anti-conservative inferences. However, for CRTs two choices for single imputation are group mean imputation and cluster mean imputation.[8] In the first case, missing outcomes in each intervention group are replaced by the mean outcome calculated using complete records pooled across clusters of that group. This approach reduces the variability among the clusters means and, therefore, gives inflated Type I error.[8] In cluster mean imputation, missing outcomes in each cluster are replaced by the mean outcome calculated using complete records of that cluster. This approach has been suggested as a good approach for handling missing outcomes by Taljaard et al.[8] They showed that cluster mean imputation gives Type I error close to nominal level under MCAR, using adjusted t-test with balanced CRTs. However, under MAR or CDM, adjusted t-test with cluster mean imputation may not be valid. We note that, with balanced CRTs, the cluster-level t-test and the adjusted t-test are identical with cluster mean imputation since after imputation the cluster sizes become constant and the cluster means remain unchanged by the imputation. Consequently, our later results for the validity of cluster level t-test can also be applied to infer the validity of results after using cluster mean imputation. One additional problem with cluster mean imputation is that it distorts the estimates of between-cluster variability and within-cluster variability, which often are of interest.

4.3 MI

MI, first proposed by Rubin,[16] is a method of filling in the missing outcomes multiple times by simulating from an appropriate model. The aim of imputing multiple times is to allow for the uncertainty about the missing outcomes due to the fact that the imputed values are sampled draws for the missing outcomes. A sequence of Q imputed datasets is obtained by replacing each missing outcome by a set of $Q \geq 2$ imputed values that are simulated from an appropriate distribution or model. Each of the Q datasets is then analysed as a completed dataset using a

standard method. The results from the Q imputed datasets are then combined using Rubin's rules.[16] The combined inference is based on a t-distribution with DF given by

$$\nu = (Q-1)\left(1+\frac{Q}{Q+1}\frac{W_{\mathrm{MI}}}{B_{\mathrm{MI}}}\right)^2 \tag{2}$$

where B_{MI} is the between-imputation variance and W_{MI} is the average within-imputation variance. This formula for DF is derived under the assumption that the complete data DF, ν_{com}, is infinite.[17]

In CRTs, ν_{com} is usually small as it is based on the number of clusters in each intervention group rather than the number of individuals. For unadjusted cluster-level analysis and individual-level baseline covariate adjusted cluster-level analysis, ν_{com} is calculated as k_1+k_2-2 for statistical inference using cluster-level t-test[4] and adjusted t-test.[8] An adjustment is made to the ν_{com} to adjust for cluster-level baseline covariates using cluster-level analysis. In this case, we reduce the complete data DF from $\nu_{\mathrm{com}}=k_1+k_2-2$ to $\nu_{\mathrm{com}}=k_1+k_2-2-p$, where p is the number of parameters corresponding to the cluster-level baseline covariates in the first-stage regression model.[4]

When ν_{com} is small and there is a modest proportion of missing data, the repeated-imputation DF, ν (given in equation (2)), for reference t-distribution can be much higher than ν_{com}, which is not appropriate.[17] In such a situation, a more appropriate DF, ν_{adj}, proposed by Barnard and Rubin,[17] is calculated as

$$\nu_{\mathrm{adj}} = \left(\frac{1}{\nu}+\frac{1}{\hat{\nu}_{\mathrm{obs}}}\right)^{-1} \leq \nu_{\mathrm{com}} \tag{3}$$

where

$$\hat{\nu}_{\mathrm{obs}} = \left(1+\frac{Q+1}{Q}\frac{B_{\mathrm{MI}}}{W_{\mathrm{MI}}}\right)^{-1}\left(\frac{\nu_{\mathrm{com}}+1}{\nu_{\mathrm{com}}+3}\right)\nu_{\mathrm{com}} \tag{4}$$

At least four different types of MI have been used in CRTs.[7] These are *standard* MI which ignores clustering, *fixed effects* MI which includes a fixed effect for each cluster in the imputation model, *random effects* MI where clustering is taken into account through random effects in the imputation model and *within-cluster* MI where standard MI is applied within each cluster. Andridge[18] showed, with balanced CRTs under MCAR and MAR missingness in a continuous outcome with a single covariate in addition to intervention indicator, that MI models that incorporate clustering using fixed effects for cluster can result in a serious overestimation of variance of group means and this overestimation is more serious for small cluster sizes and small ICCs. This overestimation of variance results in a decrease in power, which is particularly dangerous for CRTs which are often underpowered.[18] MI using random effects for cluster gave slight overestimation of variance of group means for very small values of ρ. Andridge also showed that using an MI model that ignores clustering can lead to severe underestimation of the MI variance for large values of ρ (>0.005). This underestimation of variance leads to inflated Type I error.

Taljaard et al.[8] examined the performance of MI in a simple setup considering balanced CRTs where there are no covariates except intervention indicator using standard regression imputation, which ignores clustering, and random effects MI which does account for intraclass correlation. They also considered the Approximate Bayesian Bootstrap (ABB) procedure, proposed by Rubin and Schenker,[19] as a non-parametric MI. In ABB, sampling from the posterior predictive distribution of missing data is approximated by first generating a set of plausible contributors drawn with

replacement from the observed data, and then imputed values are drawn with replacement from the possible contributors. Two possible uses of ABB in CRTs are pooled ABB and within-cluster ABB, where the set of possible contributors are sampled from all observed values across the clusters in each group or from observed values in the same cluster, respectively. They showed that none of these four MI procedures tend to yield better power compared to the power of adjusted *t*-test using no imputation and cluster mean imputation under MCAR.

We note that in the case of missing outcome under MAR for individually randomised trials, Groenwold et al.[20] showed that CRA with covariate adjustment and MI give similar estimates so long as the same set of predictors of missingness is used. It can be anticipated that similar result holds for CRTs. An obvious advantage of CRA over MI is that it is much easier to apply, and therefore in situations where they are equivalent, CRA is clearly preferable.

5 Validity of CRA of CRTs

In this section, we describe the unadjusted cluster-level analysis, baseline covariate adjusted cluster-level analysis and LMM analysis methods using complete records, and derive conditions under which they give valid inferences under the CDM assumption.

5.1 Unadjusted cluster-level analysis using complete records

The mean of the observed outcomes in the *i*th intervention group can be calculated as

$$\bar{Y}_i^{\text{obs}} = \frac{1}{k_i}\sum_{j=1}^{k_i} \bar{Y}_{ij}^{\text{obs}}$$

where $\bar{Y}_{ij}^{\text{obs}} = \left(1/\sum_{l=1}^{m_{ij}} R_{ijl}\right)\sum_{l=1}^{m_{ij}} R_{ijl}Y_{ijl}$ is the observed mean of (ij) th cluster. The estimate of intervention effect is given by

$$\hat{\theta}^{\text{obs}} = \bar{Y}_1^{\text{obs}} - \bar{Y}_2^{\text{obs}} \tag{5}$$

In Appendix 1, we show that

$$\text{E}\left(\hat{\theta}^{\text{obs}}\right) = \mu_1 - \mu_2 + \beta_1(\mu_{x11} - \mu_x) - \beta_2(\mu_{x21} - \mu_x) \tag{6}$$

and

$$\text{Var}\left(\hat{\theta}^{\text{obs}}\right) = \sum_{i=1}^{2}\frac{1}{k_i}\left(\beta_i^2\sigma_{\bar{x}_{i1}}^2 + \sigma_b^2 + \frac{\sigma_w^2}{\eta_i}\right) \tag{7}$$

where μ_{xi1} is true mean of the baseline covariate X in the *i*th intervention group among those individuals with observed outcomes, $\sigma_{\bar{x}_{i1}}^2$ is the variance of the cluster-specific means of X among those with observed outcomes and $1/\eta_i = \text{E}\left(1/\sum_l R_{ijl}\right)$. From equation (6), it follows that the unadjusted cluster-level analysis using CRA will be unbiased if

$$\beta_1(\mu_{x11} - \mu_x) = \beta_2(\mu_{x21} - \mu_x), \text{ or equivalently }, \frac{\beta_1}{\beta_2} = \frac{\mu_{x21} - \mu_x}{\mu_{x11} - \mu_x} \tag{8}$$

A sufficient condition for equation (8) to hold is that $\beta_1 = \beta_2$ (i.e. there is no interaction between baseline covariate and intervention group in the outcome model) and that the missingness

mechanisms are the same in the two intervention groups, so that $\mu_{x11} = \mu_{x21}$. It can also be seen from equation (6) that, when there is no missing data, $\mu_{x11} = \mu_{x21} = \mu_x$, and hence the unadjusted cluster-level analysis results in unbiased estimates of intervention effects even when $\beta_1 \neq \beta_2$.

5.2 Adjusted cluster-level analysis using complete records

Recall that the first step of the adjusted cluster-level analysis involves fitting a regression model for Y with X as covariate, but ignoring the intervention indicator and clustering of the data. The residual $\hat{\epsilon}_{ijl}$ is then given by

$$\hat{\epsilon}_{ijl} = Y_{ijl} - \hat{Y}_{ijl}$$

where $\hat{Y}_{ijl} = \gamma + \lambda X_{ijl}$ is the predicted outcome for the (ijl)th individual based on the first-stage model fit. The mean of the observed residuals of the ith group is given by

$$\bar{\hat{\epsilon}}_i^{\text{obs}} = \frac{1}{k_i} \sum_{j=1}^{k_i} \bar{\hat{\epsilon}}_{ij}^{\text{obs}}$$

where $\bar{\hat{\epsilon}}_{ij}^{\text{obs}} = 1/\left(\sum_{l=1}^{m_{ij}} R_{ijl}\right) \sum_{l=1}^{m_{ij}} R_{ijl}\hat{\epsilon}_{ijl}$ is the mean of observed residuals of the (ij)th cluster. The baseline covariate adjusted estimator of intervention effect is given by

$$\hat{\theta}_{\text{adj}}^{\text{obs}} = \bar{\hat{\epsilon}}_1^{\text{obs}} - \bar{\hat{\epsilon}}_2^{\text{obs}} \tag{9}$$

We show in Appendix 2 that

$$\text{E}\left(\hat{\theta}_{\text{adj}}^{\text{obs}}\right) = \mu_1 - \mu_2 + \beta_1(\mu_{x11} - \mu_x) - \beta_2(\mu_{x21} - \mu_x) + \lambda(\mu_{x21} - \mu_{x11}) \tag{10}$$

Hence, the estimator (9) will be unbiased if (i) $\beta_1 = \beta_2$ and $\mu_{x11} = \mu_{x21}$, or if (ii) $\lambda = \beta_1 = \beta_2$. Equation (10) is derived (see Appendix 2) assuming fixed values of γ and λ instead of their estimates. In practice, γ and λ are unknown and must be estimated by fitting the first-stage regression model for the observed outcomes. We are not worried about the estimate of the intercept parameter γ since the expression (10) is independent of γ. If λ is estimated consistently, then $\hat{\theta}_{\text{adj}}^{\text{obs}}$ will be a consistent estimator of intervention effect when in truth $\lambda = \beta_1 = \beta_2$. The estimator of λ, say $\hat{\lambda}$, is calculated using complete records and will be unbiased (and therefore consistent) if $R_{ijl} \perp\!\!\!\perp Y_{ijl} | X_{ijl}$. This is true only when the two intervention groups have the same missingness mechanisms and have the same baseline covariate effects on outcome in the outcome model. Therefore, assuming CDM, the baseline covariate adjusted cluster-level analysis is consistent only if the two intervention groups have the same covariate effects on outcome in the outcome model and the same missingness mechanisms. We also note that with no missing data $\mu_{x11} = \mu_{x21} = \mu_x$, hence, equation (10) guarantees that the adjusted cluster-level analysis, which assumes that the covariate effect on outcome is the same in both groups, is unbiased, regardless of whether the covariate effect is the same in the intervention groups.

The variance of the estimator (9) can be written as (see Appendix 2 for derivation)

$$\text{Var}\left(\hat{\theta}_{\text{adj}}^{\text{obs}}\right) = \sum_{i=1}^{2} \frac{1}{k_i}\left((\beta_i - \lambda)^2 \sigma_{\bar{x}_{i1}}^2 + \sigma_b^2 + \frac{\sigma_w^2}{\eta_i}\right) \tag{11}$$

This shows that when $\beta_1 = \beta_2$ and the missingness mechanisms are the same in the two intervention groups, in order for the estimator (55) to have minimum variance one should replace the unknown λ by an estimate of $\beta_1 = \beta_2 = \beta$.

5.3 LMM using complete records

Let Z be the intervention indicator which is zero for control group and is one for intervention group. When it is assumed that the two intervention groups have the same covariate effects on outcome, we fit a LMM with fixed effects of X and Z, and a random effect for cluster. Then the estimate of the coefficient of Z will be the estimated intervention effect accounting for X.

If one thinks that the baseline covariate effects on outcome could be different in the two intervention groups and there are missing outcome values, an interaction of X and Z must be included in the model. This implies that the intervention effect varies with X. Then the estimate of the intervention effect at the mean value of X is an estimate of the average intervention effect. Let X^* denote the empirically centred variable $X - \bar{X}$, where $\bar{X}$ is the mean of X calculated using data from all individuals. If the baseline covariate effects on outcome are assumed to be different in the two groups, we fit a LMM, using complete records, with fixed effects of X^*, Z and their interaction, and a random effect for cluster. The estimate of the coefficient of Z will then be the estimated average intervention effect. One may need to account for the centring step in the variance estimation. We will investigate in the simulations whether ignoring this has any negative impact on CI coverage.

In the general theory of LMM, the variances of the fixed effects parameter estimates, which are calculated based on their asymptotic distributions, are known to be underestimated for small sample sizes.[21] In this paper, we used quantiles from t-distribution with DF $k_1 + k_2 - 2$ rather than the quantiles form the standard normal distribution to construct the confidence interval for the intervention effect, as this has been used in other papers for individual-level analysis using mixed models for CRTs.[22,23]

6 Simulation study

A simulation study was conducted to investigate the performance of unadjusted cluster-level analysis, baseline covariate adjusted cluster-level analysis and LMM using CRA under baseline CDM in outcomes. We also investigated whether there is any gain using MI over CRA. The average estimate of intervention effect, its average estimated SE and coverage probability were calculated and compared. We considered balanced CRTs, where the two intervention groups have equal number of clusters ($k_i = k$) and constant cluster size ($m_{ij} = m$).

6.1 Data generation and analysis

For each individual in the study a single covariate value X was generated independently as $X \sim N(0, 1)$. Since $\sigma_x^2 = 1$, we can write the coefficient of X in equation (1) as $\beta_i = \tau_i \sigma_y$, where σ_y^2 is the total variance of Y within each intervention group and τ_i is the correlation coefficient between Y and X in intervention group i. We fixed $\sigma_y^2 = 100$, $\alpha_1 = 20$ and $\alpha_2 = 25$. Then the outcome Y was generated using the model

$$Y_{ijl} = \alpha_i + \tau_i \sigma_y X_{ijl} + \delta_{ij} + \epsilon_{ijl}$$

where $\delta_{ij} \sim N(0, \rho\sigma_y^2)$ and $\epsilon_{ijl} \sim N(0, (1-\tau_i^2-\rho)\sigma_y^2)$. We chose the cluster size $m=30$ for each cluster. Parameters that were varied in generating the data include the number of clusters in each group, $k=(5,10,20,30)$ and the unconditional ICC, $\rho=(0.001, 0.05, 0.1)$. The missing data indicators R_{ijl} under CDM assumption were generated, independently for each individual, according to a logistic regression model

$$\text{logit}\left(R_{ijl}=0|\mathbf{Y}_{ij}, \mathbf{X}_{ij}\right) = \phi_{i0} + \phi_{i1} X_{ijl}$$

The intercept ϕ_{i0} and slope ϕ_{i1} were chosen so that $\text{E}_{jl}\left(R_{ijl}\right) = p_i$, where p_i is the desired proportion of observed values in intervention group i. The degree of correlation between missingness and baseline covariate depends on the value of ϕ_{i1}. We used $\phi_{11}=\phi_{21}=1$, which gives the odds ratio for having a missing outcome (Y) is 2.72 associated with a one unit increase in the covariate (X) value. Missing data indicators were then imposed to each generated complete data to get the incomplete data.

Four possible scenarios were considered:

(1) $\phi_{10}=\phi_{20}=-1$ and $\tau_1=\tau_2=0.5$: missingness mechanism is the same between the intervention groups and there is no interaction between intervention group and baseline covariate in the outcome model.
(2) $\phi_{10}=-1$, $\phi_{20}=0.5$ and $\tau_1=\tau_2=0.5$: missingness mechanism is different between the intervention groups and there is no interaction between intervention group and baseline covariate in the outcome model.
(3) $\phi_{10}=\phi_{20}=-1$ and $\tau_1=0.4$, $\tau_2=0.6$: missingness mechanism is the same between the intervention groups and there is an interaction between intervention group and baseline covariate in the outcome model.
(4) $\phi_{10}=-1$, $\phi_{20}=0.5$ and $\tau_1=0.4$, $\tau_2=0.6$: missingness mechanism is different between the intervention groups and there is an interaction between intervention group and baseline covariate in the outcome model.

In the first and third scenarios, there was 30% missing outcomes in both the intervention groups. In the second and fourth scenarios, there was 30% missing outcomes in the control group and 60% missing outcomes in the intervention group. Each generated incomplete dataset was then analysed using unadjusted cluster-level analysis, baseline covariate adjusted cluster-level analysis and LMM using complete records. We included the interaction between intervention and covariate into the LMM in the third and fourth scenarios, where the two intervention groups have different covariate effects on outcome in the data-generating model for outcome.

The R package jomo[24] was used to multiply impute each generated incomplete dataset using MI with number of imputations 20. A random intercept LMM was used as the imputation model so that the imputation model was correctly specified. We used 200 burn-in iterations and 10 iterations between two successive draws after examining, respectively, the convergence of the posterior distributions of the parameters estimates of the imputation model and the plots of their autocorrelation functions. The completed datasets were then analysed using LMM. An interaction between intervention and baseline covariate was included in both the imputation model and the analysis model when the two intervention groups have different covariate effects on outcome in the data-generating model. We always used restricted maximum likelihood estimation method to fit the LMM. The Wald t-test with adjusted DF, given in equation (3), with $\nu_{\text{com}}=2(k-1)$ was used to test the null hypothesis of intervention effect. We had maximum 50 convergence warnings in 10,000 simulations when LMM was fitted using the R package lme4.[25]

6.2 Results

Empirical average estimates of intervention effect, average estimated SEs and coverage probabilities of nominal 95% confidence interval over 10,000 simulation runs for each of the four scenarios are presented in Tables 1 to 4, respectively.

When the missingness mechanism is the same between the intervention groups and there is no interaction between intervention and baseline covariate in the outcome model, both the unadjusted and adjusted cluster-level analyses gave unbiased estimates of intervention effect with coverage probabilities very close to the nominal level (see Table 1). However, these two methods gave biased estimates of intervention effect if the two intervention groups had either different missingness mechanisms or there was an interaction between intervention and covariate in the outcome model or both (see Tables 2 to 4). In scenario 2, (two-stage) adjusted cluster-level analysis was very slightly downwardly biased (see Table 2). Under scenario 2, the two intervention groups have the same covariate effects ($\beta_1 = \beta_2$) but the missingness mechanism is different between the intervention groups, implying $\mu_{x11} \neq \mu_{x21}$. However, although $R_{ijl} \perp\!\!\!\perp Y_{ijl} | X_{ijl}, Z_i$, $R_{ijl} \not\perp\!\!\!\perp Y_{ijl} | X_{ijl}$, where Z_i is the intervention indicator. Therefore, the estimate of regression coefficient (λ) of the first-stage analysis using CRA was biased as the regression model was fitted without considering Z_i, the intervention indicator. Consequently, the second-stage analysis gave slightly biased estimates of intervention effect. These results support our derived conditions explained in Sections 5.1 and 5.2, respectively, for unadjusted and adjusted cluster-level analyses to be unbiased using CRA, where we showed that these two methods are unbiased only if the missingness mechanism is the same between the intervention groups and there is no interaction between intervention and baseline covariate in the data-generating model for the outcome. These results also imply that cluster mean imputation, as proposed by Taljaard et al.[8]

Table 1. Simulation results-missingness mechanism is the same between the intervention groups and there is no interaction between intervention and baseline covariate in the data-generating model for outcome. Empirical average estimates of intervention effect, average estimated SEs and coverage probabilities of nominal 95% confidence interval over 10,000 simulation runs for unadjusted cluster-level analysis (CL(unadj)), baseline covariate adjusted cluster-level analysis (CL(adj)) and linear mixed model (LMM), using CRA and MI. Monte Carlo errors for average estimates and average estimated SEs are all less than 0.023 and 0.016, respectively. The true value of the intervention effect is 5.

		Average Estimate				Average estimated SE				Coverage (%)			
ρ	k	CL(unadj)	CL(adj)	LMM	MI	CL(unadj)	CL(adj)	LMM	MI	CL(unadj)	CL(adj)	LMM	MI
0.1	5	4.98	4.99	4.99	4.98	2.31	2.21	2.23	2.19	95.2	95.1	95.2	96.3
	10	5.01	4.98	5.00	4.99	1.66	1.59	1.60	1.59	95.1	95.3	95.3	95.5
	20	4.99	4.99	4.99	4.99	1.18	1.14	1.14	1.14	94.9	95.0	94.9	94.8
	30	5.01	5.00	5.01	5.01	0.97	0.93	0.93	0.93	95.0	95.0	94.9	95.0
0.05	5	5.00	4.98	5.00	5.00	1.88	1.76	1.78	1.76	95.2	95.1	95.6	96.2
	10	5.01	5.00	5.01	5.01	1.35	1.28	1.28	1.26	95.1	95.2	95.1	95.4
	20	5.01	5.00	5.01	5.01	0.96	0.91	0.91	0.90	95.0	95.0	95.1	95.0
	30	4.99	4.99	4.99	4.99	0.79	0.75	0.74	0.74	95.0	95.0	95.0	95.0
0.001	5	4.98	4.98	4.99	4.99	1.34	1.18	1.31	1.35	95.2	95.1	96.2	99.6
	10	5.01	5.00	5.01	5.01	0.96	0.85	0.90	0.93	95.1	95.1	96.8	97.8
	20	4.99	4.99	5.00	5.00	0.69	0.61	0.63	0.64	94.8	94.9	96.2	96.7
	30	5.00	5.00	5.00	5.00	0.56	0.50	0.51	0.52	95.1	95.3	96.2	96.8

Table 2. Simulation results-missingness mechanism is different between the intervention groups and there is no interaction between intervention and baseline covariate in the data-generating model for outcome. Empirical average estimates of intervention effect, average estimated SEs and coverage probabilities of nominal 95% confidence interval over 10,000 simulation runs for unadjusted cluster-level analysis (CL(unadj)), baseline covariate adjusted cluster-level analysis (CL(adj)) and linear mixed model (LMM), using CRA and MI. Monte Carlo errors for average estimates and average estimated SEs are all less than 0.025 and 0.017, respectively. The true value of the intervention effect is 5.

		Average Estimate				Average estimated SE				Coverage (%)			
ρ	k	CL(unadj)	CL(adj)	LMM	MI	CL(unadj)	CL(Adj)	LMM	MI	CL(unadj)	CL(Adj)	LMM	MI
0.1	5	3.83	4.94	5.01	5.01	2.44	2.32	2.34	2.28	93.2	95.1	95.2	97.0
	10	3.81	4.94	5.03	5.03	1.76	1.67	1.68	1.66	89.9	95.4	95.2	95.5
	20	3.78	4.91	5.00	4.99	1.25	1.19	1.19	1.19	84.2	94.9	94.8	94.8
	30	3.79	4.93	5.01	5.01	1.02	0.98	0.98	0.98	79.1	95.4	95.3	95.4
0.05	5	3.77	4.90	4.98	4.98	2.04	1.90	1.94	1.92	91.7	94.9	95.7	98.3
	10	3.78	4.90	5.00	4.99	1.48	1.38	1.38	1.36	87.5	95.0	95.0	95.8
	20	3.76	4.92	4.98	4.98	1.05	0.98	0.98	0.97	79.4	95.2	95.1	95.1
	30	3.77	4.92	4.99	4.99	0.86	0.80	0.80	0.80	70.7	94.8	94.6	94.7
0.001	5	3.77	4.89	5.00	5.00	1.58	1.39	1.54	1.60	89.4	95.1	98.3	99.7
	10	3.76	4.89	4.99	4.98	1.14	1.01	1.06	1.10	82.1	95.0	97.3	98.5
	20	3.78	4.91	5.00	5.00	0.81	0.72	0.74	0.76	68.8	95.2	96.4	97.3
	30	3.78	4.92	5.00	5.00	0.66	0.59	0.60	0.61	56.1	94.9	95.8	96.5

Table 3. Simulation results-missingness mechanism is the same between the intervention groups and there is an interaction between intervention and baseline covariate in the data-generating model for outcome. Empirical average estimates of intervention effect, average estimated SEs and coverage probabilities of nominal 95% confidence interval over 10,000 simulation runs for unadjusted cluster-level analysis (CL(unadj)), baseline covariate adjusted cluster-level analysis (CL(adj)) and linear mixed model (LMM), using CRA and MI. Monte Carlo errors for average estimates and average estimated SEs are all less than 0.024 and 0.016, respectively. The true value of the intervention effect is 5.

		Average Estimate				Average estimated SE				Coverage (%)			
ρ	k	CL(unadj)	CL(adj)	LMM	MI	CL(unadj)	CL(Adj)	LMM	MI	CL(unadj)	CL(Adj)	LMM	MI
0.1	5	4.46	4.44	4.97	4.97	2.31	2.22	2.25	2.22	94.3	94.3	95.0	96.4
	10	4.50	4.49	5.01	5.02	1.66	1.59	1.61	1.60	93.7	93.6	94.7	94.8
	20	4.48	4.48	5.00	5.00	1.19	1.14	1.15	1.15	92.5	92.6	94.9	94.9
	30	4.49	4.49	5.00	5.00	0.97	0.93	0.94	0.94	91.3	91.2	94.7	94.7
0.05	5	4.45	4.43	4.96	4.97	1.88	1.76	1.81	1.80	94.0	93.7	95.3	97.1
	10	4.51	4.49	5.01	5.01	1.36	1.28	1.30	1.29	93.7	93.4	95.0	95.5
	20	4.50	4.50	5.01	5.01	0.97	0.91	0.92	0.92	91.9	91.6	94.8	94.8
	30	4.50	4.50	5.01	5.01	0.79	0.75	0.76	0.75	90.4	89.8	94.6	94.6
0.001	5	4.48	4.46	4.99	4.99	1.34	1.18	1.35	1.39	93.4	93.5	98.1	99.4
	10	4.50	4.49	5.02	5.01	0.96	0.85	0.93	0.96	92.3	91.6	96.9	97.9
	20	4.49	4.49	5.00	5.00	0.69	0.61	0.65	0.66	88.9	87.2	96.3	96.8
	30	4.48	4.48	4.99	4.99	0.56	0.50	0.52	0.54	84.9	81.6	95.6	96.3

Table 4. Simulation results-missingness mechanism is different between the intervention groups and there is an interaction between intervention and baseline covariate in the data-generating model for outcome. Empirical average estimates of intervention effect, average estimated SEs and coverage probabilities of nominal 95% confidence interval over 10,000 simulation runs using unadjusted cluster-level analysis (CL(unadj)), baseline covariate adjusted cluster-level analysis (CL(Adj)) and linear mixed model (LMM), using CRA and MI. Monte Carlo errors for average estimates and average estimated SEs are all less than 0.025 and 0.018, respectively. The true value of the intervention effect is 5.

		Average Estimate				Average estimated SE				Coverage (%)			
ρ	k	CL(unadj)	CL(adj)	LMM	MI	CL(unadj)	CL(Adj)	LMM	MI	CL(unadj)	CL(Adj)	LMM	MI
0.1	5	3.02	4.09	5.00	5.00	2.44	2.31	2.42	2.37	89.0	93.4	95.7	98.1
	10	3.03	4.10	5.01	5.01	1.76	1.67	1.73	1.71	82.0	93.5	95.8	96.3
	20	3.03	4.11	5.01	5.01	1.25	1.19	1.23	1.23	66.6	88.8	95.6	95.6
	30	3.03	4.11	5.01	5.02	1.02	0.97	1.01	1.01	52.8	85.9	95.2	95.2
0.05	5	3.02	4.10	5.01	5.01	2.05	1.89	2.06	2.04	87.0	93.9	96.5	99.0
	10	3.02	4.10	5.01	5.01	1.47	1.36	1.45	1.44	75.9	90.4	95.7	96.7
	20	3.01	4.08	4.98	4.98	1.05	0.98	1.03	1.03	55.3	84.9	95.8	95.9
	30	3.02	4.10	5.01	5.00	0.86	0.80	0.84	0.84	38.0	81.1	95.6	95.7
0.001	5	3.02	4.07	4.99	4.99	1.57	1.37	1.69	1.75	80.4	91.1	98.5	99.8
	10	3.03	4.10	5.00	5.00	1.13	0.99	1.17	1.21	63.0	87.6	97.6	98.7
	20	3.02	4.10	5.00	5.00	0.81	0.71	0.81	0.84	33.4	77.7	97.0	97.7

(continued)

Table 5. Comparison between the complete data DF (ν_{com}) and the average estimates of adjusted DF (ν_{adj}), over 10,000 simulation runs, used by MI, when the two intervention groups have different missingness mechanisms and different covariate effects on outcome in the data-generating model for outcome (scenario 4). The last two columns show the upper 2.5% points of the t-distribution with ν_{com} and ν_{adj} DF, respectively.

ρ	k	ν_{com}	ν_{adj}	$t_{\nu_{com}}(0.025)$	$t_{\nu_{adj}}(0.025)$
0.1	5	8	4.58	2.31	2.64
	10	18	11.72	2.10	2.18
	20	38	25.71	2.02	2.06
	30	58	38.74	2.00	2.02
0.05	5	8	3.92	2.31	2.80
	10	18	9.64	2.10	2.24
	20	38	20.61	2.02	2.08
	30	58	30.18	2.00	2.04
0.001	5	8	3.12	2.31	3.11
	10	18	7.12	2.10	2.36
	20	38	13.73	2.02	2.14
	30	58	19.01	2.00	2.09

DF: degrees of freedom.

(described in Section 4.2), is not valid under CDM assumption unless the two intervention groups have the same missingness mechanisms and there is no interaction between intervention and baseline covariate in the outcome model. The bias in average intervention effect estimates could be in either direction. But, in this paper, we always have downward bias in the reported intervention effect

estimates. This is because we considered a positive correlation between baseline covariate and outcome in the data generation process, and a positive association between baseline covariate and probability of missingness in outcomes. As a result, a large value of outcome has higher chance of being missing compared to a low value of outcome. In our simulations the degree of bias was high if the two intervention groups had different covariate effects on outcome and it goes up if, in addition, the two intervention groups have different missingness mechanisms (see Tables 3 and 4). LMM and MI gave unbiased estimates of intervention effect under all the four considered scenarios, provided that an interaction of intervention and baseline covariate was included in the model to allow for different covariate effects on outcome in the two intervention groups (scenario 3 and 4).

The LMM and MI had similar empirical average estimated SEs of the intervention effect estimates. The LMM gave coverage probabilities close to nominal level except for very small ρ and small k, where it showed slightly overcoverage. However, while LMM with ν_{com} gave good coverage, MI using ν_{adj} gave overcoverage, and this can be attributed to it using a smaller DF. The average estimates of ν_{adj}, used by MI, over 10,000 simulations runs and ν_{com} for scenario 4 are presented in Table 5. Results showed that the estimates of ν_{adj} are smaller compared to ν_{com}.

7 Discussion and conclusion

In this paper, we aimed to investigate the validity of the unadjusted and adjusted cluster-level analyses, and LMM for analysing CRTs, where the outcomes are continuous and only outcomes are missing under CDM assumption. We used CRA and MI for handling the missing outcomes. The contributions of the paper can be summarised as follows:

First, we found that both the unadjusted and adjusted cluster-level analyses are in general biased using CRA unless there is no interaction between intervention and baseline covariate in the data-generating model for outcome and the missingness mechanism is the same between the interventions groups, which is arguably unlikely to hold in practice. Cluster-level analysis is used by many researchers to analyse CRTs because of its simplicity. We therefore caution researchers that these methods may commonly give biased inferences in CRTs with missing outcomes. However, we note that these two methods are unbiased with full data, even when there is an interaction between baseline covariate and intervention in the true data-generating model for outcome.

Second, cluster mean imputation has been previously recommended as a valid approach for handling missing outcomes in CRTs. We found that cluster mean imputation gave invalid inferences under CDM assumption unless missingness mechanism is the same between the intervention groups and there is no interaction between intervention and baseline covariate in the data-generating model for outcome.

Third, the LMM using CRA gave unbiased estimates of intervention effect regardless of whether missingness mechanisms are the same or are different between the intervention groups and whether there is an interaction between intervention and baseline covariate in the data-generating model for the outcome, provided that an interaction between intervention and baseline covariate was included in the model when such interaction exists in truth.

Finally, we compared the results of LMM using CRA with the results of MI. As expected, we found that MI gave unbiased intervention effects estimates regardless of whether missingness mechanisms are the same or are different in the two intervention groups and whether there is an interaction between intervention and baseline covariate. The LMM and MI had similar empirical SEs of the estimates of intervention effects. However, MI using adjusted DF estimates gave overcoverage for the nominal 95% confidence interval. This is due to underestimation of adjusted DF used by MI compared to complete data DF. Groenwold et al.[20] showed that there is little to be

gained by using MI over LMM in the absence of auxiliary variables. Moreover, when missingness is confined to outcomes, LMMs fitted using maximum likelihood are fully efficient and valid under MAR.

Throughout this paper, we have assumed CDM mechanism in a continuous outcome, which is an example of MAR as our baseline covariate was fully observed. In practice, we cannot identify on the basis of the observed data which missingness assumption is appropriate.[14,26] Therefore, sensitivity analyses should be performed[26] (Ch. 10) to explore whether our inferences are robust to the primary working assumption regarding the missingness mechanism. Furthermore, we focused on studies with only one individual-level covariate; the methods described can be extended for more than one covariate.

In conclusion, in the absence of auxiliary variables, LMM using complete records can be recommended as the primary analysis approach for CRTs with missing outcomes if one is willing to make baseline CDM assumption for outcomes.

Acknowledgements

The authors thank the anonymous reviewers for their comments and constructive suggestions which led to an improvement over the earlier version of the manuscript.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: A Hossain was supported by the Economic and Social Research Council (ESRC), UK, via Bloomsbury Doctoral Training Centre (ES/J5000021/1). K Diaz-Ordaz was funded by Medical Research Council (MRC) career development award in Biostatistics (MR/L011964/1). J W Bartlett was supported by MRC fellowship (MR/K02180X/1) while a member of the Department of Medical Statistics, London School of Hygiene & Tropical Medicine (LSHTM).

References

1. Murray DM and Blitstein JL. Methods to reduce the impact of interclass correlation in group-randomised trials. *Eval Rev* 2003; **27**: 79–103.
2. Donner A and Klar N. *Design and analysis of cluster randomization trials in health research*. London: Arnold, 2000.
3. Murray DM. *Design and analysis of group-randomized trials*. New York: Oxford University Press, 1998.
4. Hayes RJ and Moulton LH. *Cluster randomised trials*. London: CRC Press, Taylor & Francis Group, 2009.
5. Wood AM, White IR and Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clin Trials* 2004; **1**: 368–376.
6. Sterner JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009; **339**: 157–160.
7. Diaz-Ordaz K, Kenward MG, Cohen A, et al. Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. *Clin Trials* 2014; **11**: 590–600.
8. Taljaard M, Donner A and Klar N. Imputation strategies for missing continuous outcomes in cluster randomized trials. *Biometr J* 2008; **50**: 329–345.
9. Gail MH, Tan WY and Piantadosi S. Tests for no treatment effect in randomised clinical trials. *Biometrika* 1988; **75**: 57–64.
10. Hernandez AV, Steyerberg EW and Habbema JD. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *J Clin Epidemiol* 2004; **57**: 454–460.
11. Hubbard AE, Ahern J, Fleischer NL, et al. To GEE or not to GEE: comparing population average and mixed models

for estimating the associations between neighborhood risk factors and health. *Epidemiology* 2010; **21**: 467–474.
12. Little RJA and Rubin DB. *Statistical analysis with missing data*, 2nd ed. New Jersey: John Wiley & Sons, 2002.
13. Rubin DB. Inference and missing data. *Biometrika* 1976; **63**: 581–592.
14. White IR and Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med* 2010; **29**: 2920–2931.
15. Bartlett JW, Carpenter JR, Tilling K, et al. Improving upon the efficiency of complete case analysis when covariates are MNAR. *Biostatistics* 2014; **15**: 719–730.
16. Rubin DB. *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons, 1987.
17. Barnard J and Rubin DB. Small-sample degrees of freedom with multiple imputation. *Biometrika* 1999; **86**: 948–955.
18. Andridge RR. Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometr J* 2011; **53**: 57–74.
19. Rubin DB and Schenker N. Multiple imputation for interval estimation from simple random samples with ignorable nonresponses. *J Am Stat Assoc* 1986; **81**: 366–374.
20. Groenwold RH, Donders AR, Roes KC, et al. Dealing with missing outcome data in randomized trials and observational studies. *Am J Epidemiol* 2012; **175**: 210–217.
21. Kenward MG and Roger JH. Small sample inference for fixed effects from restricted maximum likelihoods. *Biometrics* 1997; **53**: 983–997.
22. Ukoumunne OC, Carlin JB and Gulliford MC. A simulation study of odds ratio estimation for binary outcomes from cluster randomized trials. *Stat Med* 2007; **26**: 3415–3428.
23. Ma J, Thabane L, Kaczorowski J, et al. Comparison of Bayesian and classical methods in the analysis of cluster randomized controlled trials with a binary outcome: the Community Hypertension Assessment Trial (CHAT). *BMC Med Res Methodol* 2009; **9**: 37.
24. Quartagno M and Carpenter J. jomo: A package for multilevel joint modelling multiple imputation, September 2015, http://CRAN.R-project.org/package=jomo.
25. Bates D, Maechler M, Bolker B, et al. Fitting linear mixed-effects models using lme4. *J Stat Softw* 2015; **67**: 1–48.
26. Carpenter JR and Kenward MG. *Multiple imputations and its applications*. United Kingdom: John Wiley & Sons, 2013.

Appendix 1. Unadjusted cluster-level analysis using complete records

The mean of the observed outcomes in a particular cluster can be written as

$$\begin{aligned}
\bar{Y}_{ij}^{\text{obs}} &= \frac{1}{\sum_{l}^{m_{ij}} R_{ijl}} \sum_{l=1}^{m_{ij}} R_{ijl} Y_{ijl} \\
&= \frac{1}{\sum_{l}^{m_{ij}} R_{ijl}} \sum_{l=1}^{m_{ij}} R_{ijl} \left(\alpha_i + \beta_i X_{ijl} + \delta_{ij} + \epsilon_{ijl}\right) \\
&= \alpha_i + \beta_i \frac{1}{\sum_{l}^{m_{ij}} R_{ijl}} \sum_{l=1}^{m_{ij}} R_{ijl} X_{ijl} + \delta_{ij} + \frac{1}{\sum_{l}^{m_{ij}} R_{ijl}} \sum_{l=1}^{m_{ij}} R_{ijl} \epsilon_{ijl} \\
&= \alpha_i + \beta_i \bar{X}_{ij}^{\text{obs}} + \delta_{ij} + \frac{1}{\sum_{l}^{m_{ij}} R_{ijl}} \sum_{l=1}^{m_{ij}} R_{ijl} \epsilon_{ijl}
\end{aligned}$$

where $\bar{X}_{ij}^{\text{obs}} = \left(1/\sum_l R_{ijl}\right) \sum_{l=1}^{m_{ij}} R_{ijl} X_{ijl}$ is the observed mean of the baseline covariate X in the (ij)th cluster. The expected value of $\bar{X}_{ij}^{\text{obs}}$ across the clusters in the ith intervention group will be the true mean of X among those individuals with observed outcomes. Let μ_{xi1} denote the true mean of the baseline covariate X in the ith intervention group among those individuals with observed outcomes. Then

$$\text{E}\left(\bar{Y}_{ij}^{\text{obs}}\right) = \alpha_i + \beta_i \mu_{xi1} + \text{E}\left(\frac{1}{\sum_{l}^{m_{ij}} R_{ijl}} \sum_{l=1}^{m_{ij}} R_{ijl} \epsilon_{ijl}\right)$$

Let $\mathbf{R_{ij}} = (R_{ij1}, R_{ij2}, \ldots, R_{ijm_{ij}})$ be the vector of missing data indicators for the (ij)th cluster. Then

$$\begin{aligned}
\mathrm{E}\left(\frac{1}{\sum_l R_{ijl}}\sum_{l=1}^{m_{ij}} R_{ijl}\epsilon_{ijl}\right) &= \mathrm{E}\left[\mathrm{E}\left(\frac{1}{\sum_l R_{ijl}}\sum_{l=1}^{m_{ij}} R_{ijl}\epsilon_{ijl}|\mathbf{R_{ij}}\right)\right] \\
&= \mathrm{E}\left[\frac{1}{\sum_l R_{ijl}}\sum_{l=1}^{m_{ij}} R_{ijl}\mathrm{E}\left(\epsilon_{ijl}|\mathbf{R_{ij}}\right)\right] \\
&= 0 \qquad (12)
\end{aligned}$$

since ϵ_{ijl}'s are independent of R_{ijl}'s and $\mathrm{E}(\epsilon_{ijl}) = 0$. Therefore, we have

$$\mathrm{E}\left(\bar{Y}_{ij}^{\mathrm{obs}}\right) = \alpha_i + \beta_i\mu_{xi1}$$

The variance of $\bar{Y}_{ij}$ can be written as

$$\begin{aligned}
\mathrm{Var}\left(\bar{Y}_{ij}^{\mathrm{obs}}\right) &= \beta_i^2\mathrm{Var}\left(\bar{X}_{ij}^{\mathrm{obs}}\right) + \sigma_b^2 + \mathrm{Var}\left(\frac{1}{\sum_l R_{ijl}}\sum_{l=1}^{m_{ij}} R_{ijl}\epsilon_{ijl}\right) \\
&= \beta_i^2\sigma_{\bar{x}_{i1}}^2 + \sigma_b^2 + \mathrm{Var}\left(\frac{1}{\sum_l R_{ijl}}\sum_{l=1}^{m_{ij}} R_{ijl}\epsilon_{ijl}\right)
\end{aligned}$$

where $\sigma_{\bar{x}_{i1}}^2$ is the variance of the cluster-specific means of X among those with observed outcomes.

Now

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{\sum_l^{m_{ij}} R_{ijl}}\sum_{l=1}^{m_{ij}} R_{ijl}\epsilon_{ijl}\right) &= \mathrm{Var}\left[\mathrm{E}\left(\frac{1}{\sum_l R_{ijl}}\sum_{l=1}^{m_{iij}} R_{ijl}\epsilon_{ijl}|\mathbf{R}_{ij}\right)\right] \\
&\quad + \mathrm{E}\left[\mathrm{Var}\left(\frac{1}{\sum_l R_{ijl}}\sum_{l=1}^{m_{ij}} R_{ijl}\epsilon_{ijl}|\mathbf{R}_{ij}\right)\right] \\
&= 0 + \mathrm{E}\left[\frac{1}{\left(\sum_l R_{ijl}\right)^2}\sum_{l=1}^{m_{ij}} R_{ijl}\mathrm{Var}\left(\epsilon_{ijl}|\mathbf{R}_{ij}\right)\right], \text{ using equation (12)} \\
&= \sigma_w^2\mathrm{E}\left(\frac{1}{\sum_l R_{ijl}}\right) \\
&= \frac{\sigma_w^2}{\eta_i} \qquad (13)
\end{aligned}$$

where $\mathrm{E}\left(1/\left(\sum_l^{m_{ij}} R_{ijl}\right)\right) = 1/\eta_i$ (say). Therefore

$$\mathrm{Var}\left(\bar{Y}_{ij}^{\mathrm{obs}}\right) = \beta_i^2\sigma_{\bar{x}_{i1}}^2 + \sigma_b^2 + \frac{\sigma_w^2}{\eta_i}$$

The observed mean of the ith intervention group is calculated as

$$\bar{Y}_i^{\mathrm{obs}} = \frac{1}{k_i}\sum_{j=1}^{k_i} \bar{Y}_{ij}^{\mathrm{obs}}$$

Then

$$\mathrm{E}\left(\bar{Y}_i^{\mathrm{obs}}\right) = \alpha_i + \beta_i\mu_{xi1}$$

and

$$\mathrm{Var}\left(\bar{Y}_i^{\mathrm{obs}}\right) = \frac{1}{k_i}\left(\beta_i^2\sigma_{\bar{x}_{i1}}^2 + \sigma_b^2 + \frac{\sigma_w^2}{\eta_i}\right)$$

The estimator of intervention effect in unadjusted cluster-level analysis based on observed values is given by

$$\hat{\theta}^{\mathrm{obs}} = \bar{Y}_1^{\mathrm{obs}} - \bar{Y}_2^{\mathrm{obs}}$$

Then

$$\begin{aligned}\mathrm{E}\left(\hat{\theta}^{\mathrm{obs}}\right) &= (\alpha_1 + \beta_1\mu_{x11}) - (\alpha_2 + \beta_2\mu_{x21}) \\ &= (\alpha_1 + \beta_1\mu_x) - (\alpha_2 + \beta_2\mu_x) + \beta_1(\mu_{x11} - \mu_x) - \beta_2(\mu_{x21} - \mu_x) \\ &= \mu_1 - \mu_2 + \beta_1(\mu_{x11} - \mu_x) - \beta_2(\mu_{x21} - \mu_x)\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}\left(\hat{\theta}^{\mathrm{obs}}\right) &= \frac{1}{k_1}\left(\beta_1^2\sigma_{\bar{x}_{11}}^2 + \sigma_b^2 + \frac{\sigma_w^2}{\eta_1}\right) + \frac{1}{k_2}\left(\beta_2^2\sigma_{\bar{x}_{21}}^2 + \sigma_b^2 + \frac{\sigma_w^2}{\eta_2}\right) \\ &= \sum_{i=1}^{2}\frac{1}{k_i}\left(\beta_i^2\sigma_{\bar{x}_{i1}}^2 + \sigma_b^2 + \frac{\sigma_w^2}{\eta_i}\right)\end{aligned}$$

which tends to zero as (k_1, k_2) tend to infinity.

Appendix 2. Adjusted cluster-level analysis using complete records

The mean of observed residuals of a particular cluster is given by

$$\bar{\hat{\epsilon}}_{ij}^{\mathrm{obs}} = \frac{1}{\sum_l^{m_{ij}} R_{ijl}}\sum_{l=1}^{m_{ij}} R_{ijl}\hat{\epsilon}_{ijl}$$

$$= \frac{1}{\sum_{l} R_{ijl}} \sum_{l=1}^{m_{ij}} R_{ijl}\left(Y_{ijl} - \hat{Y}_{ijl}\right)$$

$$= \frac{1}{\sum_{l} R_{ijl}} \sum_{l=1}^{m_{ij}} R_{ijl}\left(\alpha_i + \beta_i X_{ijl} + \delta_{ij} + \epsilon_{ijl} - \gamma - \lambda X_{ijl}\right)$$

$$= \alpha_i + (\beta_i - \lambda)\frac{1}{\sum_{l} R_{ijl}} \sum_{l=1}^{m_{ij}} R_{ijl} X_{ijl} + \delta_{ij} + \frac{1}{\sum_{l} R_{ijl}} \sum_{l=1}^{m_{ij}} R_{ijl}\epsilon_{ijl} - \gamma$$

$$= \alpha_i + (\beta_i - \lambda)\bar{X}_{ij}^{\text{obs}} + \delta_{ij} + \frac{1}{\sum_{l} R_{ijl}} \sum_{l=1}^{m_{ij}} R_{ijl}\epsilon_{ijl} - \gamma$$

Then

$$\mathrm{E}\left(\bar{\hat{\epsilon}}_{ij}^{\text{obs}}\right) = \alpha_i + (\beta_i - \lambda)\mu_{xi1} - \gamma$$

and

$$\mathrm{Var}\left(\bar{\hat{\epsilon}}_{ij}^{\text{obs}}\right) = (\beta_i - \lambda)^2 \sigma^2_{\bar{x}_{i1}} + \sigma_b^2 + \frac{\sigma_w^2}{\eta_i}$$

using the results (12) and (13). The mean of observed residuals of the ith intervention group can be written as

$$\bar{\hat{\epsilon}}_{i}^{\text{obs}} = \frac{1}{k_i} \sum_{j=1}^{k_i} \bar{\hat{\epsilon}}_{ij}^{\text{obs}}$$

Then

$$\mathrm{E}\left(\bar{\hat{\epsilon}}_{i}^{\text{obs}}\right) = \alpha_i + (\beta_i - \lambda)\mu_{xi1} - \gamma$$

and

$$\mathrm{Var}\left(\bar{\hat{\epsilon}}_{i}^{\text{obs}}\right) = \frac{1}{k_i}\left((\beta_i - \lambda)^2 \sigma^2_{\bar{x}_{i1}} + \sigma_b^2 + \frac{\sigma_w^2}{\eta_i}\right)$$

The baseline covariate adjusted estimator of intervention effect, based on observed values, is given by

$$\hat{\theta}_{\text{adj}}^{\text{obs}} = \bar{\hat{\epsilon}}_{1}^{\text{obs}} - \bar{\hat{\epsilon}}_{2}^{\text{obs}}$$

Then

$$\begin{aligned}
\mathrm{E}\left(\hat{\theta}_{\text{adj}}^{\text{obs}}\right) &= (\alpha_1 + (\beta_1 - \lambda)\mu_{x11} - \gamma) - (\alpha_2 + (\beta_2 - \lambda)\mu_{x21} - \gamma) \\
&= (\alpha_1 + \beta_1\mu_x) - (\alpha_2 + \beta_2\mu_x) + \beta_1(\mu_{x11} - \mu_x) - \beta_2(\mu_{x21} - \mu_x) + \lambda(\mu_{x21} - \mu_{x11}) \\
&= \mu_1 - \mu_2 + \beta_1(\mu_{x11} - \mu_x) - \beta_2(\mu_{x21} - \mu_x) + \lambda(\mu_{x21} - \mu_{x11})
\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}\left(\hat{\theta}_{\mathrm{adj}}^{\mathrm{obs}}\right) &= \frac{1}{k_1}\left((\beta_1-\lambda)^2\sigma^2_{\bar{x}_{11}}+\sigma_b^2+\frac{\sigma_w^2}{\eta_1}\right)+\frac{1}{k_2}\left((\beta_2-\lambda)^2\sigma^2_{\bar{x}_{21}}+\sigma_b^2+\frac{\sigma_w^2}{\eta_2}\right)\\ &= \sum_{i=1}^{2}\frac{1}{k_i}\left((\beta_i-\lambda)^2\sigma^2_{\bar{x}_{i1}}+\sigma_b^2+\frac{\sigma_w^2}{\eta_i}\right)\end{aligned}$$

which tends to zero as (k_1, k_2) tend to infinity.

Part III

Binary Outcomes

Chapter 6

Review of Analysis Methods with Full Data and Missing Data

In this chapter, we discuss the terminology, define the necessary notations used in this part of the thesis, and review the literature on handling missing binary outcomes in CRTs. Section 6.1 describes the two broad approaches to the analysis of binary outcomes in CRTs. In Section 6.2, we review the literature on handling missing binary outcomes in CRTs, and identify the research questions. In the following chapter, we will present our research on missing binary outcomes in CRTs.

6.1 Analysis with full data

In this section, we briefly review the two broad approaches to the analysis of binary outcomes in CRTs with full data. These approaches are cluster-level analysis and individual-level analysis. Let Y_{ijl} be a binary outcome of interest for the lth ($l = 1, 2, \ldots, m_{ij}$) individual in the jth ($j = 1, 2, \ldots, k_i$) cluster of the ith ($i = 0, 1$) intervention group, where $i = 0$ corresponds to control group and $i = 1$ corresponds to active intervention group. Also let X_{ijl} be an individual-level baseline covariate value for lth individual in the (ij)th cluster. Note that these methods can be extended to the case of multiple baseline covariates, some of which are at the individual-level and some are at the cluster-level. For convenience, we assume that both control and intervention groups have the same number of clusters ($k_i = k$) and constant cluster size across the groups ($m_{ij} = m$).

6.1.1 Cluster-level analysis

This approach is conceptually very simple and can be explained as a two-stage approach. There are two different types of cluster-level analysis. These are unadjusted cluster-level analysis and (baseline covariate) adjusted cluster-level analysis. For binary outcomes, risk difference (RD) or risk ratio (RR) is usually estimated as a measure of intervention effect using cluster-level analysis in CRTs. The cluster-specific proportion of success is usually used as the summary measure for each cluster. In unadjusted cluster-level analysis, RD is estimated as the difference between the means of the cluster-specific proportions of success in the two interventions groups, and RR is estimated as the ratio

of the means of the cluster-specific proportions of success in the two interventions groups. In the second stage, a test of the hypothesis RD = 0 or log(RR) = 0 is performed using an appropriate statistical method. The most popular one is the standard $t-$test for two independent samples. The reasons for using this test is that the cluster-specific summary measures are statistically independent, which is a consequence of the clusters being independent from each other.

In an adjusted cluster-level analysis, an individual-level regression analysis of the outcome of interest is carried out at the first stage of analysis ignoring the clustering of the data, which incorporates all covariates into the regression model except intervention indicator [5, 7]. A standard logistic regression model is usually fitted for binary outcomes [5]. Then the observed and predicted numbers of success are compared by computing a residual for each cluster. In the case of no intervention effect, the residuals should be similar on average in the two intervention groups. In the case of calculating adjusted RD, the residual, known as difference-residual, is calculated for each cluster as the difference between the observed and predicted proportions of success. Then adjusted RD is estimated as the difference between the means of the cluster-specific difference residuals of the two interventions groups. In the case of calculating adjusted RR, the residual, known as ratio-residual, is calculated for each cluster as the ratio of the observed number of success to the predicted number of success. Then adjusted RR is estimated as the ratio of the means of the cluster-specific ratio residuals of the two intervention groups.

6.1.2 Individual-level analysis

In individual-level analysis, a regression model is fitted to the individual-level outcome which allows us to estimate the fixed effect coefficients corresponding to intervention indicator and other covariates, if any. There are two types of models for CRTs. These are cluster-specific (CS) (also known as conditional) models and population averaged (PA) (also known as marginal) models. The CS models estimate the effect of intervention on outcome while cluster random effect is held constant, known as CS intervention effect. In contrast, the PA models estimate the effect of intervention on outcome as averaged over all clusters, i.e, over the range of random effects. For a linear model, both the CS and PA models estimate the same population parameter. However, for non-linear models, this is not necessarily the case. For binary outcomes in CRTs, the most commonly used CS model is random-effects logistic regression (RELR) model and the most commonly used PA model is generalised estimation equations (GEE). Both RELR and GEE are extensions of the standard logistic regression models modified to allow for correlation between the outcomes of individuals in the same cluster.

Random-effects logistic regression (RELR) model takes into account of between-cluster variability by incorporating cluster-specific random effects, which are almost always assumed to be normally distributed, into the logistic regression. These models are fitted by maximising the likelihood function numerically, because the likelihood function and its derivative can not be derived analytically as this involves an integral over the distribution of the random effects. Numerical integration methods are used to approximate the integral and so approximate the likelihood function.

Generalised estimating equations (GEE) take into account the correlation among the outcomes of the same cluster using a working correlation matrix. In CRTs, it is usual to assume that the correlation matrix is exchangeable, since outcomes on individuals in different clusters are uncorrelated, while outcomes on individuals in the same cluster are equally correlated. In GEE, the sandwich standard error estimator is typically used to estimate the standard error of the parameter estimates. Although the sandwich standard error estimator is consistent even when the working correlation structure is specified incorrectly, it tends to be biased downwards when the number of clusters in each intervention group is small [5, 34]. Moreover, the estimate of standard error is highly variable when the number of clusters is small. It is recommended to have at least 40 clusters in the study to get reliable standard error estimates [35]. A number of methods have been proposed for dealing with the limitations of the sandwich variance estimator [34, 36]. Ukoumunne (2007) [34] suggested the following method to correct the bias for small number of clusters in each intervention group. Firstly, the downward bias of the sandwich standard error estimator is adjusted by multiplying it by $\sqrt{k/(k-1)}$, where k is the number of clusters in each intervention group. Secondly, the increased small sample variability of the sandwich standard error estimator is accounted for by constructing the confidence interval for intervention effect based on the quantiles from a $t-$distribution rather than quantiles from standard normal distribution.

6.2 Analysis with missing outcomes

A number of recent studies [25–28] had investigated how to handle missing binary outcomes in CRTs under the assumption of CDM. However, these studies simulated datasets in ways which arguably do not correspond to how data arise in CRTs, raising doubts about their conclusions.

Ma *et al.* [25] examined within-cluster MI, fixed effects MI and MMI for missing binary outcomes under CDM mechanism in CRTs, using RELR and GEE. They showed that all these strategies give quite similar results for low percentages of missing data or for small value of ICC. With high percentage of missing data, they found that within-cluster MI underestimated the variance of the intervention effect which result in inflated Type I error rate. However, the simulation study was based on a real dataset, so the conclusions to other design settings may be limited. It is therefore difficult to draw conclusions from their results about the performance of GEE and RELR with different MI strategies under CDM mechanism.

In two subsequent studies, Ma *et al.* [26, 27] compared the performance of GEE and RELR with missing binary outcomes under CDM mechanism using CRA, standard MI and within-cluster MI. They concluded that GEE using CRA performs well in terms of bias when the percentage of missing outcomes is low. In contrast, they concluded that RELR using CRA does not perform well. However, in the case of missing outcomes under MAR for individually randomised trials, Groenwold *et al.* [15] showed that CRA with covariate adjustment and MI give similar estimates as long as the same functional form of the same set of predictors of missingness are used. It can be anticipated that

a similar result holds for CRTs. Moreover, in the case of missing continuous outcomes under CDM in CRTs, we showed in the published paper in Chapter 5 that LMM using CRA adjusted for covariates such that the CDM assumption holds give unbiased estimates of intervention effect. Similar conclusion can be anticipated in the case of binary outcomes in CRTs using RELR and GEE.

These two studies by Ma *et al.* [26, 27] also concluded that GEE performs well when using standard MI and the variance inflation factor (VIF) is less than 3; and when using within-cluster MI and VIF ≥ 3 with cluster size is at least 50. In contrast, they concluded that RELR does not perform well using either standard MI or within-cluster MI. Their simulation study showed that standardised bias for RELR with full data were much higher than those obtained by standard MI or within-cluster MI. However, we expect zero bias or possibly small finite sample bias with full data. The reasons for contradictory conclusions by Ma *et al.* [26, 27] are because they generated the data in such a way that they knew what the true PA log(OR) was, but after fitting RELR, they compared estimates of CS log(OR) to the true PA log(OR), which are expected to be different due to non-collapsibility. In addition, in the data generating mechanism used in these studies [26, 27], the baseline covariate was generated independently of the outcome, which in general is not a plausible assumption. It is therefore difficult to draw conclusions about what would happen in CRTs where the baseline covariates are related to the outcome.

Caille *et al.* [28] compared different MI strategies through a simulation study for handing missing binary outcomes in CRTs assuming CDM. They showed that GEE using unadjusted CRA and using adjusted (for covariates) CRA are biased for estimating in-

tervention effects. However, as we stated earlier, it is expected that GEE using CRA adjusted for covariates give unbiased estimates of intervention effect if the CDM assumption holds. In their simulation study, individual-level continuous outcomes were generated at first using a LMM which included intervention indicator and a cluster random effect for each cluster, but without covariates. Each continuous outcome was then dichotomised to obtain a binary outcome. Then, baseline covariates were generated dependent on the continuous outcomes. So it appears the data generation mechanism used would mean that baseline covariates were associated with intervention group, which is not possible (in expectation) due to randomisation. In addition, as the authors noted, they compared estimates of CS ORs to the true PA ORs, which is expected to differ even with full data due to non-collapsibility. It is therefore difficult to draw general conclusions from their results about the methods' performance in CRTs. Caille *et al.* [28] also concluded that that MMI with RELR and single-level MI with standard logistic regression give better inference for intervention effect compared to CRA in terms of bias, efficiency and coverage. However, their data generation mechanism does not correspond to how data arise in CRTs. It is therefore again difficult to draw general conclusions from their results about the MI strategies' performance in CRTs.

All of these previous studies [25–28] considered only individual-level analysis and estimated odds ratio (OR) as a measure of intervention effect. The risk difference (RD) or risk ratio (RR) may be of interest as measures of intervention effect, and have a number of advantages over OR [37]. For example, they are arguably easier to understand, and they are 'collapsible', i.e., the population marginal and conditional (on covariates or cluster effects or both) values are identical in the absence of confounding. Cluster-level analysis methods can be used to analyse CRTs where RD or RR is estimated

as a measure of intervention effect [5], and these analyses can also incorporate adjustment for baseline covariates. These methods have the advantage of being simple to apply compared to the individual-level analysis methods. To date the performance of cluster-level analysis approaches with incompletely-observed binary outcomes has not been investigated.

In the research paper contained in Chapter 7, we will investigate the validity of estimating RD and RR as measures of intervention effect using unadjusted and adjusted cluster-level analysis methods when binary outcomes are missing under a CDM mechanism. We will also investigate the validity of RELR and GEE considering the limitations of previous studies [25–28], which we described earlier in this Section.

Chapter 7

Research Paper II

Title: Missing binary outcomes under covariate dependent missingness in cluster randomised trials.

Author(s): Anower Hossain, Karla DiazOrdaz and Jonathan W. Bartlett.

Journal Name: Statistics in Medicine.

Type of publication: Research paper.

Stage of Publication: Volume 36, Issue 19, August 2017, Pages 3092-3109.

URL: `http://onlinelibrary.wiley.com/doi/10.1002/sim.7334/epdf`

DOI: 10.1002/sim.7334

Academic peer-reviewed: Yes.

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
www.lshtm.ac.uk

Registry
T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Anower Hossain (AH)
Principal Supervisor	Karla Diaz-Ordaz
Thesis Title	Missing data in cluster randomised trials

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?	Statistics in Medicine		
When was the work published?	June 2017		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	NA		
Have you retained the copyright for the work?*	Yes	Was the work subject to academic peer review?	Yes

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	AH identified the research questions, designed and run the simulation study, and interpreted the results with supervision from the supervisors. AH wrote the initial draft of the manuscript and then revised it based on the feedback received from the supervisors. All authors read and approved the final manuscript.

Student Signature: Anower Hossain **Date:** 29.08.17

Supervisor Signature: [signature] **Date:** 29/08/17

Summary of Research Paper II

Title: Missing binary outcomes under covariate dependent missingness in cluster randomised trials.

This paper investigates the validity of estimating RD and RR as measures of intervention effect using unadjusted and adjusted cluster-level analysis methods when binary outcomes are missing under a CDM mechanism. In addition, it investigates the validity of individual-level analysis approaches considering the limitations of previous studies.

We show analytically and through simulations that cluster-level analyses for estimating RD using complete records are in general biased. For estimating RR, cluster-level analyses using complete records are valid if the true data generating model has log link, and the intervention groups have the same missingness mechanism and the same covariate effect in the outcome model. In contrast, MMI followed by cluster-level analyses give unbiased estimates of RD and RR regardless of whether missingness mechanisms were the same or different between the intervention groups and whether there is an interaction between intervention and baseline covariate in the outcome model, provided that this interaction is included in the imputation model when required.

In the case of individual-level analysis, as long as both MMI and CRA use the same functional form of the same set of baseline covariates, RELR or GEE using CRA adjusted for covariates such that the CDM assumption holds can be recommended as the primary analysis approach for CRTs with missing binary outcomes if one is willing to make the CDM assumption for outcomes.

Statistics
in Medicine

Research Article

Received 19 August 2016, Accepted 20 April 2017 Published online 29 May 2017 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7334

Missing binary outcomes under covariate-dependent missingness in cluster randomised trials

Anower Hossain,[a,b*†] Karla DiazOrdaz[a] and Jonathan W. Bartlett[c]

Missing outcomes are a commonly occurring problem for cluster randomised trials, which can lead to biased and inefficient inference if ignored or handled inappropriately. Two approaches for analysing such trials are cluster-level analysis and individual-level analysis. In this study, we assessed the performance of unadjusted cluster-level analysis, baseline covariate-adjusted cluster-level analysis, random effects logistic regression and generalised estimating equations when binary outcomes are missing under a baseline covariate-dependent missingness mechanism. Missing outcomes were handled using complete records analysis and multilevel multiple imputation. We analytically show that cluster-level analyses for estimating risk ratio using complete records are valid if the true data generating model has log link and the intervention groups have the same missingness mechanism and the same covariate effect in the outcome model. We performed a simulation study considering four different scenarios, depending on whether the missingness mechanisms are the same or different between the intervention groups and whether there is an interaction between intervention group and baseline covariate in the outcome model. On the basis of the simulation study and analytical results, we give guidance on the conditions under which each approach is valid. © 2017 The Authors. *Statistics in Medicine* **Published by John Wiley & Sons Ltd.**

Keywords: cluster randomised trials; missing binary outcome; baseline covariate-dependent missingness; complete records analysis; multiple imputation

1. Introduction

Cluster randomised trials (CRTs), also known as group randomised trials, are increasingly being used to evaluate the effectiveness of interventions in health services research [1,2]. The unit of randomisation for such trials are identifiable clusters of individuals such as medical practices, schools or entire communities. However, individual-level outcomes of interest are observed within each cluster. One important feature of CRTs is that the outcomes of individuals within the same cluster are more likely to be similar to each other than those from different clusters, which is usually quantified by the intraclass correlation coefficient (ICC, denoted as ρ). Although typically in primary care and health research the value of ICC is small ($0.001 < \rho < 0.05$) [3], it can lead to substantial variance inflation factors and should not be ignored [2,4]. This is because ignoring the dependence of the outcomes of individuals within the clusters will underestimate the variance of the intervention effect estimates and consequently give inflated type I error rates [5]. It is well known that the power and precision of CRTs are lower compared with trials that individually randomise the same number of units [2]. However, in practice, CRTs have several advantages including that the nature of the intervention itself may dictate its application at the cluster level, less risk of intervention contamination and administrative convenience [6]. These advantages are sometimes judged by researchers to outweigh the potential loss of statistical power and precision.

[a]*Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, U.K.*
[b]*Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka 1000, Bangladesh*
[c]*Statistical Innovation Group, AstraZeneca, Cambridge, U.K.*
[*]*Correspondence to: Anower Hossain, Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka 1000, Bangladesh.*
[†]*E-mail: anower@isrt.ac.bd*

Missing data are a commonly occurring threat to the validity and efficiency of CRTs. In a systematic review of CRTs published in English in 2011, 72% of trials had missing values either in outcomes or in covariates or in both, and only 34% of them reported how missing data had been handled [7]. Dealing with missing data in CRTs is complicated because of the clustering of the data. In statistical analysis, if there are missing values, an assumption must be made about the relationship between the probability of data being missing and the underlying values of the variables involved in the analysis. The mechanisms that caused the data to be missing can be classified into three broad categories. Data are missing completely at random (MCAR) if the probability of missingness is independent of the observed and unobserved data. MCAR is generally a very restrictive assumption and is unlikely to hold in many studies. A more plausible assumption is missing at random (MAR) where, conditioning on the observed data, the probability of missingness is independent of the unobserved data. Missing not at random is the situation where the probability of missingness depends on both the observed and unobserved data. In CRTs, an assumption regarding missing outcomes that is sometimes plausible is that missingness depends on baseline covariates, but conditioning on these baseline covariates, not on the outcome itself. We refer to this as covariate-dependent missingness (CDM). This is an example of MAR when baseline covariates are fully observed. In this paper, we will consider the case of a binary outcome that is partially observed and assume that all baseline covariates are fully observed.

Two approaches for analysing CRTs are cluster-level analyses, which derive summary statistics for each cluster, and individual-level analyses, which use the data for each individual in each cluster [6]. Complete records analysis (CRA) and multiple imputation (MI) (described in Section 3) are the most commonly used methods for handling missing data. A number of recent studies have investigated how to handle missing binary outcomes in CRTs under the assumption of CDM [8–11]. However, as we describe in detail in Section 3, these previous studies simulated datasets in ways that arguably do not correspond to how data arise in CRTs raising doubt about their conclusions.

In the case of missing outcome under MAR for individually randomised trials, Groenwold *et al.* [12] showed that CRA with covariate adjustment and MI give similar estimates as long as the same covariates and same functional form are used. It can be anticipated that a similar result holds for CRTs. In the case of missing continuous outcomes in CRTs, Hossain *et al.* [13] showed that there is no gain in terms of bias or efficiency of the estimates using MI over CRA adjusted for covariates, where both approaches used the same covariates with the same functional form, and the same modelling assumptions. Therefore in situations where they are equivalent, CRA is clearly preferable.

All of these previous studies [8–11] considered only individual-level analysis and estimated odds ratio (OR) as a measure of intervention effect. The risk difference (RD) or risk ratio (RR) may be of interest as measures of intervention effect and have a number of advantages over OR [14]. For example, they are arguably easier to understand, and they are 'collapsible', that is, the population marginal and conditional (on covariates or cluster effects or both) values are identical. Cluster-level analysis methods can be used to analyse CRTs where RD or RR is estimated as a measure of intervention effect [6], and these analyses can also incorporate adjustment for baseline covariates. These methods have the advantage of being simple to apply compared with the individual-level analysis methods. To date, the performance of cluster-level analysis approaches with incompletely observed binary outcomes has not been investigated.

The aim of this paper is twofold. The first is to investigate the validity of estimating RD and RR as measures of intervention effect using unadjusted and adjusted cluster-level analysis methods when binary outcomes are missing under a CDM mechanism. The second is to investigate the validity of individual-level analysis approaches considering the limitations of previous studies [8–11], which we describe in Section 3. CRA and MI are used to handle the missing outcomes.

This paper is organised as follows. We begin in Section 2 by giving a brief review of the approaches to the analysis of binary outcome in CRTs with full data. Section 3 describes methods of handling missing data in CRTs. In Section 4, we investigate the validity of CRA of CRTs under CDM assumption for missing binary outcomes. In Section 5, we report the results of a simulation study to investigate the performance of our considered methods. Section 6 presents an example of application of our results to an actual CRT. We conclude in Section 7 with some discussion.

2. Analysis of CRTs with full data

We begin by describing the two broad approaches to the analysis of CRTs in the absence of missing data. These two approaches are cluster-level analysis and individual-level analysis. Let Y_{ijl} be a binary outcome

of interest for the lth ($l = 1, 2, \ldots, m_{ij}$) individual in the jth ($j = 1, 2, \ldots, k_i$) cluster of the ith ($i = 0, 1$) intervention group, where $i = 0$ corresponds to control group and $i = 1$ corresponds to intervention group. For convenience, we assume that both control and intervention groups have the same number of clusters ($k_i = k$) and constant cluster size across the groups ($m_{ij} = m$). Also let X_{ijl} be an individual-level baseline covariate value for lth individual in the (ij)th cluster. Note that these methods can be extended to the case of multiple baseline covariates, some of which are individual level and some are cluster level.

In the case of a continuous outcome, it is common to assume that the expectation of the outcome is linearly dependent on the covariate and intervention indicator. However, this assumption is not very plausible in the case of a binary outcome. Two commonly used alternatives in the case of binary outcome are to assume a log or logit link between the mean of the outcome and the linear predictor.

In the case of a log link, each binary Y_{ijl} is assumed to be generated by

$$\pi_{ijl} = \exp(\beta_0 + \beta_1 i + f_i(X_{ijl}) + \delta_{ij}), \tag{1}$$

where β_0 is a constant, β_1 is the true intervention effect, $f_i(X_{ijl})$ is a function of baseline covariate X_{ijl} in the ith intervention group, δ_{ij} is the (ij)th cluster effect with mean 0 and $\pi_{ijl} = P\left(Y_{ijl} = 1 | \delta_{ij}, X_{ijl}\right)$. On the other hand, assuming a logit link for the true data generating model, we have

$$\pi_{ijl} = \text{expit}\left(\beta_0 + \beta_1 i + f_i(X_{ijl}) + \delta_{ij}\right), \tag{2}$$

where $\text{expit}(t) = \exp(t)/(1 + \exp(t))$.

2.1. *Cluster-level analysis*

This approach is conceptually very simple and can be explained as a two-stage process. Two different ways of doing cluster-level analysis are unadjusted cluster-level analysis and (baseline covariate) adjusted cluster-level analysis. For binary outcomes, RD or RR is usually estimated as a measure of intervention effect in cluster-level analysis [6].

2.1.1. Unadjusted cluster-level analysis (CL_U). In the first stage of analysis, a relevant summary measure of outcomes is obtained for each cluster. For binary outcomes, the cluster-level proportion of success is usually used as the summary measure for each cluster. Let p_{ij} be the observed proportion of success in the (ij)th cluster. Then RD is estimated as

$$\widehat{\text{RD}}_{\text{unadj}} = \bar{p}_1 - \bar{p}_0,$$

where $\bar{p}_i$ is the mean of the cluster-specific proportions of success in the ith intervention group. In the second stage, a test of the hypothesis RD = 0 is performed using an appropriate statistical method. The most popular one is the standard t-test for two independent samples with degrees of freedom (DF) $2k - 2$. The reason for using this test is that the cluster-specific summary measures are statistically independent, which is a consequence of the clusters being independent of each other.

On the basis of the first stage cluster-level summary measures, RR is estimated as

$$\widehat{\text{RR}}_{\text{unadj}} = \frac{\bar{p}_1}{\bar{p}_0}.$$

Then, in the second stage, a test of the hypothesis $\log(\text{RR}) = 0$ is performed using t-test with DF $2k - 2$, where $\hat{\text{V}}\left(\log(\widehat{\text{RR}}_{\text{unadj}})\right)$ can be calculated as [6]

$$\hat{\text{V}}\left(\log(\widehat{\text{RR}}_{\text{unadj}})\right) \approx \frac{s_0^2}{k\bar{p}_0^2} + \frac{s_1^2}{k\bar{p}_1^2} \qquad \text{with} \qquad s_i^2 = \frac{\sum_{j=1}^{k}\left(p_{ij} - \bar{p}_i\right)^2}{k - 1}.$$

It can be shown that, with full data, $\widehat{\text{RD}}_{\text{unadj}}$ is unbiased for RD, and $\widehat{\text{RR}}_{\text{unadj}}$ is consistent (and, therefore, asymptotically unbiased) for RR as $k \to \infty$ (see Appendix A in the Supporting Information).

2.1.2. Adjusted cluster-level analysis (CL_A). In CRTs, baseline covariates that may be related to the outcome of interest are often collected and incorporated into the analysis. The main purpose of adjusting for covariates is to increase the credibility of the trial findings by demonstrating that any observed intervention effect is not attributable to the possible imbalance between the intervention groups in terms of baseline covariates [15].

In an adjusted cluster-level analysis, an individual-level regression analysis of the outcome of interest is carried out at the first stage of analysis ignoring the clustering of the data, which incorporates all covariates into the regression model except intervention indicator [6, 16]. A standard logistic regression model is usually fitted for binary outcomes, which assumes that

$$\text{logit}\left(\pi_{ijl}\right) = \log\left(\frac{\pi_{ijl}}{1-\pi_{ijl}}\right) = \lambda_1 + \lambda_2 X_{ijl}. \tag{3}$$

Let N_{ij} and $\hat{N}_{ij}$ be the observed and predicted number of successes in the (ij)th cluster, respectively. After fitting model (3), $\hat{N}_{ij}$ is calculated as

$$\hat{N}_{ij} = \sum_{l=1}^{m} \hat{\pi}_{ijl} = \sum_{l=1}^{m} \text{expit}\left(\hat{\lambda}_1 + \hat{\lambda}_2 X_{ijl}\right).$$

Then the observed and predicted numbers of success are compared by computing a residual for each cluster. In the case of no intervention effect, the residuals should be similar on average in the two intervention groups.

If we want to estimate the adjusted RD, the residual, known as difference residual, for each cluster is calculated as $\epsilon_{ij}^d = (N_{ij} - \hat{N}_{ij})/m$, where the d superscript refers to difference residual. The adjusted RD is then estimated as

$$\widehat{\text{RD}}_{\text{adj}} = \bar{\epsilon}_1^d - \bar{\epsilon}_0^d,$$

where $\bar{\epsilon}_i^d$ is the mean of the difference residuals across the clusters of the ith intervention group and where $\widehat{\text{RD}}_{\text{adj}}$ can be rewritten as

$$\widehat{\text{RD}}_{\text{adj}} = \widehat{\text{RD}}_{\text{unadj}} + \frac{1}{mk}\sum_{j=1}^{k}\left(\hat{N}_{0j} - \hat{N}_{1j}\right). \tag{4}$$

Because the distribution of X (in expectation) is the same between the intervention groups as a consequence of randomisation, and the prediction from the first-stage regression model (3) depends only on X_{ijl}, $\text{E}\left(\hat{N}_{0j}\right) = \text{E}\left(\hat{N}_{1j}\right)$. Hence, from (4), $\widehat{\text{RD}}_{\text{adj}}$ is unbiased for RD because $\widehat{\text{RD}}_{\text{unadj}}$ is unbiased for RD. In the second stage, a test of hypothesis $\text{RD}_{\text{adj}} = 0$ is performed using t-test with DF $2k-2$.

If we want to estimate the adjusted RR, the residual, also known as ratio residual, for each cluster is calculated as $\epsilon_{ij}^r = N_{ij}/\hat{N}_{ij}$, where the r superscript refers to ratio residual. The adjusted RR is then estimated as

$$\widehat{\text{RR}}_{\text{adj}} = \frac{\bar{\epsilon}_1^r}{\bar{\epsilon}_0^r}, \tag{5}$$

where $\bar{\epsilon}_i^r$ is the mean of the ratio residuals across the clusters of the ith intervention group. It can be shown that, with full data, $\widehat{\text{RR}}_{\text{adj}}$ is consistent and, therefore, asymptotically unbiased (as $k \to \infty$) for true RR if (i) the true data generating model is a log link model; (ii) the functional form of the covariates is the same between the intervention groups; and (iii) the distribution of random effect is the same between the intervention groups (see Appendix B in the Supporting Information for details). In the second stage, a test of hypothesis $\log\left(\text{RR}_{\text{adj}}\right) = 0$ is performed using t-test with DF $2k-2$, where $\hat{\text{V}}\left(\log(\widehat{\text{RR}}_{\text{adj}})\right)$ can be calculated as

$$\hat{\text{V}}\left(\log(\widehat{\text{RR}}_{\text{adj}})\right) \approx \frac{s_{\epsilon 0}^2}{k\left(\bar{\epsilon}_0^r\right)^2} + \frac{s_{\epsilon 1}^2}{k\left(\bar{\epsilon}_1^r\right)^2} \qquad \text{with} \qquad s_{\epsilon i}^2 = \frac{\sum_{j=1}^{k}\left(\epsilon_{ij}^r - \bar{\epsilon}_i^r\right)^2}{k-1}.$$

2.2. *Individual-level analysis*

In individual-level analysis, a regression model is fitted to the individual-level outcome that allows us to analyse the effects of intervention and other covariates in the same model. For binary outcomes, two commonly used individual-level analysis methods are random effects logistic regression (RELR), which estimates cluster-specific (also known as conditional) intervention effects, and generalised estimation equations (GEEs), which estimate population-averaged (also known as marginal) intervention effects. Both of these approaches are extensions of the standard logistic regression models modified to allow for correlation between the outcomes of individuals in the same cluster. We also note that for both methods, one can obtain estimates of RD or RR by integrating over the fixed and random effects in the case of RELR and by integrating over the fixed effects in the case of GEE.

2.2.1. Random effects logistic regression. RELR models take into account between-cluster variability by incorporating cluster-specific random effects, which are almost always assumed to be normally distributed, into the logistic regression. These models are fitted by maximising the likelihood function numerically, because the likelihood function and its derivative cannot be derived analytically as this involves an integral over the distribution of the random effects. Numerical integration methods are used to approximate the integral and so approximate the likelihood function. It is recommended to have at least 15 clusters in each intervention group to acquire the correct size and coverage for significance tests and confidence interval [6]. Li and Redden [17] examined the performance of five denominator degrees of freedom (DDF) approximations, namely, residual DDF, containment DDF, between-within DDF, Satterthwaite DDF and Kenward–Roger DDF. They recommended to use between-within DDF approximation, which is equal to the total number of clusters in the study minus the rank of the design matrix, as it gave type I error rate close to nominal level and higher power compared with the other four methods. Ukoumunne *et al.* [18] examined the properties of t-based confidence intervals for log(OR) from CRTs using DF $2k-2$ assuming the same number of clusters in the two intervention groups. They found that the coverage rates were close to the nominal level, although this approach gave overcoverage with very small ICC (0.001). In this paper, we used the quantiles from t-distribution with DF $2k-2$ rather than quantiles from $\mathcal{N}(0,1)$ to construct the confidence interval for intervention effect.

2.2.2. Generalised estimating equations. GEEs are commonly used as a method for analysing binary outcomes in CRTs while taking into account the correlation among the outcomes of the same cluster using a working correlation matrix. In CRTs, it is usual to assume that the correlation matrix is exchangeable, because outcomes on individuals in different clusters are uncorrelated, while outcomes on individuals in the same cluster are equally correlated.

In GEE, the sandwich standard error (SE) estimator is typically used to estimate the SE of the parameter estimates. Although the sandwich SE estimator is consistent even when the working correlation structure is specified incorrectly, the sandwich SE of the regression coefficient tends to be biased downwards when the number of clusters in each intervention group is small [6, 18]. Moreover, the estimate of SE is highly variable when the number of clusters is small. It is recommended to have at least 40 clusters in the study to acquire reliable SE estimates [5]. A number of methods have been proposed for dealing with the limitations of the sandwich variance estimator [18, 19]. In this paper, we used the method proposed by Ukoumunne (2007) [18] to correct the bias for small number of clusters in each intervention group. Firstly, the downward bias of the sandwich SE estimator was adjusted by multiplying it by $\sqrt{k/(k-1)}$, where k is the number of clusters in each intervention group. Secondly, the increased small sample variability of the sandwich SE estimator was accounted for by constructing the confidence interval for intervention effect on the basis of the quantiles from a t-distribution with DF $2k-2$ rather than quantiles from $\mathcal{N}(0,1)$. However, if some baseline covariates were cluster level, the DF would be adjusted downwards as $2k-2-q$ to account for this, where q is the number of parameters corresponding to the cluster-level baseline covariates.

3. Methods of handling missing data in CRTs

Common methods for handling missing data in CRTs are CRA, single imputation and MI. In this paper, we focused on CRA and MI because they are the most commonly used methods for handling missing

data. All the analysis methods described in the previous section can be implemented using either complete records or MI. This section briefly describes these two approaches.

3.1. Complete records analysis

In CRA, often referred to as complete case analysis, only individuals with complete data on all variables in the analysis are considered. It has the advantage of being simple to apply and is usually the default method in most statistical packages. It is well known that CRA is valid if data are MCAR. CRA is also valid if, conditioning on covariates, missingness is independent of outcome and and the outcome model being fitted is correctly specified [20]. On the basis of simulations for CDM in CRTs, Ma *et al.* [9, 10] showed that GEE using CRA performs well in terms of bias when the percentage of missing outcomes is low. In contrast, they concluded that RELR using CRA does not perform well. This is because they generated the data in such a way that they knew what the true population-averaged log(OR) was, but after fitting RELR, they compared estimates of conditional (on cluster random effects and covariates) log(OR) with the true population-averaged log(OR). In addition, in the data generating mechanism used in these studies [9, 10], the baseline covariate was generated independently of the outcome, which in general is not a plausible assumption. It is therefore difficult to draw conclusions about what would happen in CRTs where the baseline covariates are related to the outcome. Caille *et al.* [11] reported through simulations that GEE using unadjusted CRA and using adjusted (for covariates) CRA are biased for estimating intervention effects. However, in their simulation study, individual-level continuous outcomes were generated at first using a linear mixed model that includes intervention indicator and a cluster random effect for each cluster, but without covariates. Each continuous outcome was then dichotomised to obtain a binary outcome. Then, baseline covariates were generated dependent on the continuous outcomes. So it appears the data generation mechanism used would mean that baseline covariates were associated with intervention group, which is not possible (in expectation) because of randomisation. In addition, as the authors noted, they compared estimates of covariate conditional ORs with the true unconditional ORs, which would be expected to differ even with full data because of non-collapsibility. It is therefore difficult to draw general conclusions from their results about the methods' performance in CRTs.

3.2. Multiple imputation

In MI, a sequence of Q imputed datasets are obtained by replacing each missing outcome by a set of $Q \geq 2$ imputed values that are simulated from an appropriate distribution or model. Imputing multiple times allows the uncertainty associated with the imputed values because the imputed values are sampled draws for the missing outcomes instead of the actual values. This uncertainty is taken into account by adding between-imputation variance to the average within-imputation variance. Each of the Q imputed datasets are analysed as a full dataset using standard methods, and the results are then combined using Rubin's rules [21]. One important feature of MI is that the imputation model and the analysis model do not have to be the same. However, in order for Rubin's rules to be valid, the imputation model needs to be compatible or congenial with the analysis model [22].

There are at least four different types of MI that have been used in CRTs [7]. These are *standard* MI, also known as *single-level* MI, that ignores clustering in the imputation model, *fixed effects* MI that includes a fixed effect for each cluster in the imputation model, *random effects* MI where clustering is taken into account through a random effect for each cluster in the imputation model and *within-cluster* MI where standard MI is applied within each cluster. From now, we refer to random effects MI as multilevel multiple imputation (MMI).

The MI inference is usually based on a t-distribution with DF given by

$$\upsilon = (Q-1)\left(1+\frac{Q}{Q+1}\frac{W}{B}\right)^2,$$

where B and W are the between-imputation variance and the average within-imputation variance, respectively. This DF is derived under the assumption that the complete data (full data) DF, υ_{com}, is infinite [23]. In CRTs, the value of υ_{com} is calculated on the basis of the number of clusters in the study rather than the number of individuals and, therefore, is usually small. In CRTs with equal number of clusters in each intervention group, υ_{com} is calculated as $2k-2$ [24]. If υ_{com} is small and there is a modest proportion of missing data, the value of υ can be much higher than υ_{com}, which is not appropriate [23]. In such

a situation, a more appropriate DF, proposed by Barnard and Rubin (1999) [23], is calculated as

$$\nu_{\text{adj}} = \left(\upsilon^{-1} + \hat{\upsilon}_{\text{obs}}^{-1}\right)^{-1} \leq \nu_{\text{com}} \quad \text{where} \quad \hat{\nu}_{\text{obs}} = \left(\frac{\nu_{\text{com}}+1}{\nu_{\text{com}}+3}\right)\nu_{\text{com}}\left(1+\frac{Q+1}{Q}\frac{B}{W}\right)^{-1}.$$

Ma *et al.* [8] examined within-cluster MI, fixed effects MI and MMI for missing binary outcomes under CDM mechanism in CRTs. They showed that all these strategies give quite similar results for low percentages of missing data or for small value of ICC. With high percentage of missing data, the within-cluster MI underestimates the variance of the intervention effect that may result in inflated type I error rate. In two subsequent studies, Ma *et al.* [9, 10] compared the performance of GEE and RELR with missing binary outcomes using standard MI and within-cluster MI. Results showed that GEE performs well when using standard MI and the variance inflation factor is less than 3 and using within-cluster MI when variance inflation factor is ≥3 and cluster size is at least 50. Ma *et al.* [10] concluded that RELR does not perform well using either standard MI or within-cluster MI. However, in the latter two studies [9, 10], as we described in Section 3.1, they compared estimates of conditional (on cluster random effects and covariates) log(OR) with the true population-averaged log(OR), and their data generation mechanisms do not correspond to how data arise in CRTs. In the first study [8], the simulation was based on a real dataset, so the conclusions to other design settings may be limited. It is therefore again difficult to draw conclusions from their results about the performance of GEE and RELR with different MI strategies under CDM mechanism. Caille *et al.* [11] compared different MI strategies through a simulation study for handing missing binary outcomes in CRTs assuming CDM, assessing bias, SE and coverage rate of the estimated intervention effect. They showed that MMI with RELR and single-level MI with standard logistic regression give better inference for intervention effect compared with CRA in terms of bias, efficiency and coverage. However, as we described in Section 3.1, their data generation mechanism does not correspond to how data arise in CRTs. It is therefore again difficult to draw general conclusions from their results about the MI strategies' performance in CRTs.

In the case of missing continuous outcome in CRTs, Andridge [24] showed that the true MI variance of group means are underestimated by single-level MI and are overestimated by fixed effects MI. She also showed that MMI is the best among these three methods and recommended its use for practitioners. DiazOrdaz *et al.* [25] showed that for bivariate outcomes, MMI gives coverage rate close to nominal level, whereas single-level MI gives low coverage and fixed effects MI gives overcoverage. In this paper, we therefore used MMI for missing binary outcome.

4. Validity of CRA of CRTs

In this section, we investigate the validity of CL_{U}, CL_{A}, RELR and GEE using complete records, when binary outcomes are missing under CDM.

In settings where the expectation of the outcome is assumed to be linearly dependent on the covariate and intervention indicator, both unadjusted and adjusted cluster-level analyses using complete records for estimating mean difference as a measure of intervention effect are unbiased in general only when the two intervention groups have the same CDM mechanism and the same covariate effect on the outcome [13]. However, as described in Section 2, the assumption of the expectation of the outcome being linear in baseline covariate and intervention indicator is not very plausible in the case of a binary outcome. Two common alternatives are to use a log or logit link between the mean of the outcome and the linear predictor.

Define a missing outcome data indicator R_{ijl} such that $R_{ijl} = 1$ if Y_{ijl} is observed and $R_{ijl} = 0$ if Y_{ijl} is missing. Then $\sum_{l=1}^{m} R_{ijl}$ is the number of complete records in the (ij)th cluster.

4.1. Cluster-level analyses for estimating RD

In unadjusted cluster-level analysis using complete records, RD is estimated as

$$\widehat{\text{RD}}^{\text{cr}}_{\text{unadj}} = \bar{p}_1^{\text{cr}} - \bar{p}_0^{\text{cr}},$$

where $\bar{p}_i^{\text{cr}}$ is the mean of the cluster-specific proportions of success, calculated using complete records, in the ith intervention group. The superscript **cr** refers to complete records.

In adjusted cluster-level analysis, recall that a logistic regression model is fitted to the data at the first stage of analysis ignoring intervention and clustering of the data. Then the observed and predicted number of successes in each cluster are compared by computing a residual for each cluster. The adjusted RD using complete records is estimated as

$$\widehat{\mathrm{RD}}^{\mathrm{cr}}_{\mathrm{adj}} = \bar{\epsilon}^{d(\mathrm{cr})}_{1} - \bar{\epsilon}^{d(\mathrm{cr})}_{0},$$

where $\bar{\epsilon}^{d(\mathrm{cr})}_{i}$ is the average of the cluster-specific difference residuals in the ith intervention group using complete records. Then $\widehat{\mathrm{RD}}^{\mathrm{cr}}_{\mathrm{adj}}$ can be written in terms of $\widehat{\mathrm{RD}}^{\mathrm{cr}}_{\mathrm{unadj}}$ as

$$\widehat{\mathrm{RD}}^{\mathrm{cr}}_{\mathrm{adj}} = \widehat{\mathrm{RD}}^{\mathrm{cr}}_{\mathrm{unadj}} + \frac{1}{k}\sum_{j=1}^{k}\left[\frac{1}{\sum_{l=1}^{m} R_{ijl}}\left(\hat{N}^{\mathrm{cr}}_{0j} - \hat{N}^{\mathrm{cr}}_{1j}\right)\right], \tag{6}$$

where $\hat{N}^{\mathrm{cr}}_{ij}$ is the predicted number of successes using complete records in the (ij)th cluster.

We aim to derive conditions under which the cluster-level analyses for RD using complete records are unbiased. To this end, we write the individual-level probabilities of success, π_{ijl}, as

$$\pi_{ijl} = \pi_i + g_i\left(X_{ijl}, \delta_{ij}\right),$$

where $g_i\left(X_{ijl}, \delta_{ij}\right)$ is a function of baseline covariate X_{ijl} and random cluster effect δ_{ij} and which determines how individual-level probabilities of success differ from group-level probability of success in each intervention group. Then it can be shown that $\widehat{\mathrm{RD}}^{\mathrm{cr}}_{\mathrm{unadj}}$ will be unbiased for true RD if and only if

$$\mathrm{E}\left(g_1\left(X_{1jl}, \delta_{1j}\right)|R_{1jl}=1\right) = \mathrm{E}\left(g_0\left(X_{0jl}, \delta_{0j}\right)|R_{0jl}=1\right), \tag{7}$$

(see Appendix C of the Supporting Information for more details). Assuming the data are generated from log link model (1) or logit link model (2) and there is an intervention effect ($\beta_1 \neq 0$) in truth, the condition (7) is not satisfied even if the two intervention groups have the same missingness mechanism and the same covariate effects in the data generating model for the outcome. Hence, $\widehat{\mathrm{RD}}^{\mathrm{cr}}_{\mathrm{unadj}}$ is biased for true RD ($\neq 0$) when the true data generating model has log link or logit link. However, under the null hypothesis of no intervention effect ($\beta_1 = 0$), if the two intervention groups have the same covariate effects and the same missingness mechanism, the condition (7) is satisfied, and hence, $\widehat{\mathrm{RD}}^{\mathrm{cr}}_{\mathrm{unadj}}$ is unbiased for true RD = 0.

Referring to equation (6), if the two intervention groups have the same missingness mechanism and the same covariate effect, then $\mathrm{E}\left(\hat{N}^{\mathrm{cr}}_{0j}\right) = \mathrm{E}\left(\hat{N}^{\mathrm{cr}}_{1j}\right)$. Hence, with $\beta_1 \neq 0$, from equation (6), we can conclude that because $\widehat{\mathrm{RD}}^{\mathrm{cr}}_{\mathrm{unadj}}$ is biased for RD ($\neq 0$) with both log and logit links for the true data generating model, $\widehat{\mathrm{RD}}^{\mathrm{cr}}_{\mathrm{adj}}$ is also biased for RD ($\neq 0$) with both log and logit links in the true data generating model. However, with $\beta_1 = 0$, since $\widehat{\mathrm{RD}}^{\mathrm{cr}}_{\mathrm{unadj}}$ is unbiased for RD = 0 with both log and logit links, when the two intervention groups have the same missingness mechanism and the same covariate effect, $\widehat{\mathrm{RD}}^{\mathrm{cr}}_{\mathrm{adj}}$ is also unbiased for RD = 0 under the same conditions. It can also be shown that the expectation of $g_i\left(X_{ijl}, \delta_{ij}\right)$ over (j, l) is zero for $i \in \{0, 1\}$ for both log and logit links in the data generating model, and hence, both $\widehat{\mathrm{RD}}_{\mathrm{unadj}}$ and $\widehat{\mathrm{RD}}_{\mathrm{adj}}$ are unbiased for true RD with full data.

4.2. *Cluster-level analyses for estimating RR*

In both unadjusted and adjusted cluster-level analyses, RR is estimated using complete records as, respectively,

$$\widehat{\mathrm{RR}}^{\mathrm{cr}}_{\mathrm{unadj}} = \frac{\bar{p}^{\mathrm{cr}}_{1}}{\bar{p}^{\mathrm{cr}}_{0}} \quad \text{and} \quad \widehat{\mathrm{RR}}^{\mathrm{cr}}_{\mathrm{adj}} = \frac{\bar{\epsilon}^{r(\mathrm{cr})}_{1}}{\bar{\epsilon}^{r(\mathrm{cr})}_{0}}, \tag{8}$$

where $\bar{\epsilon}^{r(\mathrm{cr})}_{i}$ is the average of the ratio residuals in the ith intervention group using complete records.

We aim to establish conditions under which the cluster-level analyses for RR using complete records are consistent. To this end, we write π_{ijl} as

$$\pi_{ijl} = \pi_i\, h_i\left(X_{ijl}, \delta_{ij}\right),$$

where $h_i\left(X_{ijl}, \delta_{ij}\right)$ is a function of baseline covariate X_{ijl} and random cluster effect δ_{ij} and which determines how individual-level probabilities of success differ from group-level probability of success. Then it can be shown that $\widehat{\mathrm{RR}}^{\mathrm{cr}}_{\mathrm{unadj}}$ will be consistent for true RR if only if

$$\frac{\mathrm{E}\left(h_1\left(X_{1jl}, \delta_{1j}\right)|R_{1jl}=1\right)}{\mathrm{E}\left(h_0\left(X_{0jl}, \delta_{0j}\right)|R_{0jl}=1\right)}=1, \tag{9}$$

(see Appendix D of the Supporting Information for more details). Assuming the data are generated from log link model (1), the condition (9) is satisfied if the two intervention groups have the same missingness mechanism and the same covariate effects, and hence, $\widehat{\mathrm{RR}}^{\mathrm{cr}}_{\mathrm{unadj}}$ is consistent (and, therefore, asymptotically unbiased) for true RR.

On the other hand, assuming the data are generated from logit link model (2) with $\beta_1 \neq 0$, the condition (9) is not satisfied even if the two intervention groups have the same missingness mechanism and the same covariate effects. Hence, $\widehat{\mathrm{RR}}^{\mathrm{cr}}_{\mathrm{unadj}}$ is not consistent for true RR ($\neq 1$). However, under the null hypothesis of no intervention effect ($\beta_1 = 0$), if the two intervention group have the same missingness mechanism and the same covariate effect, the condition (9) is satisfied, and hence, $\widehat{\mathrm{RR}}^{\mathrm{cr}}_{\mathrm{unadj}}$ is consistent for true RR = 1.

In Appendix E of the Supporting Information, we show that $\widehat{\mathrm{RR}}^{\mathrm{cr}}_{\mathrm{adj}}$ is consistent and, therefore, asymptotically unbiased (as $k \to \infty$) for true RR if (i) the true data generating model is a log link model, (ii) the functional form of the covariates in the outcome model is the same between the intervention groups, (iii) the missingness mechanism is the same between the intervention groups and (iv) the distribution of random effects is the same between the intervention groups. If the data are generated from logit link model (2) with $\beta_1 \neq 0$, $\widehat{\mathrm{RR}}^{\mathrm{cr}}_{\mathrm{adj}}$ is not consistent for true RR ($\neq 1$). However, under the null hypothesis of no intervention effect ($\beta_1 = 0$), $\widehat{\mathrm{RR}}^{\mathrm{cr}}_{\mathrm{adj}}$ is consistent (as $k \to \infty$) for true RR ($= 1$) if (i) the true data generating model is a logit link model, (ii) the functional form of the covariates is the same between the intervention groups, (iii) the missingness mechanism is the same between the intervention groups and (iv) the distribution of random effects is the same between the intervention groups.

4.3. RELR and GEE using complete records

For individually randomised trials, it is well known that likelihood-based CRA is valid under MAR, if missingness is only in the outcome and all predictors of missingness are included in the model as covariates [20]. So it is anticipated that RELR using CRA will give consistent estimates of intervention effect, if the covariate X, which is associated with the missingness, is included in the model and the model is correctly specified. We also expect that GEE using CRA adjusted for covariate X that is associated with the missingness in outcomes will give consistent estimates of intervention effect.

When it is assumed that the two intervention groups have the same covariate effects on outcome, we fit RELR with fixed effects of intervention indicator and covariate and a random effect for cluster, while we fit GEE with intervention indicator and covariate assuming exchangeable correlation for the outcomes of the same cluster. If it is assumed that the baseline covariate effect on outcome could be different in the two intervention groups, an interaction between intervention and covariate must be included in the model. This implies that the intervention effect varies with level of covariate values. In those scenarios where an interaction is present, we will target the intervention effect at the mean value of the covariate. Let X^* denote the empirically centred covariate $X - \bar{X}$, where $\bar{X}$ is the mean of X using data from all individuals. Then, we fit RELR with fixed effects of intervention indicator, X^* and their interaction, and a random effect for cluster, while we fit GEE including the intervention indicator, X^* and their interaction, and assuming an exchangeable correlation for the outcomes of the same cluster. One may need to account for the centring step in the variance estimation. We will investigate in the simulation whether ignoring this has any negative impact on confidence interval coverage.

5. Simulation study

A simulation study was conducted to assess the performance of $\mathrm{CL_U}$, $\mathrm{CL_A}$, RELR and GEE under CDM mechanism. CRA and MMI were used to handle the missing data. The average estimate of intervention

effect, its average estimated SE and coverage rates were calculated for each of the methods and compared with each other. We considered balanced CRTs, where the two intervention groups have the same number of clusters and constant cluster size (before missing outcomes were introduced), and a single continuous individual-level baseline covariate.

5.1. Data generation

Data were generated using the model in equation (2) with a logit link, as described in Section 2, with $f_i(X_{ijl}) = \beta_{2(i)}X_{ijl}$, where $\beta_{2(i)}$ is the effect of covariate of X in the ith intervention group. For each individual in the study, a value of X_{ijl} was generated using the model

$$X_{ijl} = \alpha_{ij} + u_{ijl},$$

where α_{ij} is the (ij)th cluster effect on X and u_{ijl} is the individual-level error on X. We assumed that $\alpha_{ij} \sim \mathcal{N}\left(\mu_x, \sigma_\alpha^2\right)$ independently of $u_{ijl} \sim \mathcal{N}\left(0, \sigma_u^2\right)$, where μ_x is the mean of X, σ_α^2 and σ_u^2 are the between-cluster and within-cluster variance of X, respectively. The total variance of X can be written as $\sigma_x^2 = \sigma_\alpha^2 + \sigma_u^2$, and thus, the ICC of X is $\rho_x = \sigma_\alpha^2/\sigma_x^2$. Then, we generated logit(π_{ijl}) for each individual in the study using model (2) assuming $\delta_{ij} \sim \mathcal{N}\left(0, \sigma_b^2\right)$. Finally, Y_{ijl} was generated as Bernoulli random variable with parameter π_{ijl}.

Once the complete datasets (full data) were generated, we introduced missing outcomes by generating a missing outcome data indicator R_{ijl} (defined in Section (4)), independently for each individual, under CDM mechanism according to a logistic regression model

$$\text{logit}(R_{ijl} = 0 | \boldsymbol{Y}_{ij}, \boldsymbol{X_{ij}}) = \psi_i + \phi_i X_{ijl}, \tag{10}$$

where $\boldsymbol{Y}_{ij}$ and $\boldsymbol{X}_{ij}$ are the vectors of outcome and covariate values, respectively, of the (ij)th cluster. The constants ψ_i and ϕ_i were chosen such that the ith intervention group had the desired proportion of observed outcomes. The value of ϕ_i in equation (10) represents the degree of association between the missingness and the covariate X in the ith intervention group. In this study, we assumed the same covariate effects for the probability of having a missing outcome in the two intervention groups and thus set $\phi_0 = \phi_1 = 1$ in equation (10) corresponding to the OR of having a missing outcome of 2.72 for a 1 unit change in X.

We investigated four scenarios, varying whether the baseline covariate effects on outcome and the missingness mechanisms were the same in the two intervention groups. For generating X_{ijl}, we chose $\mu_x = 0, \sigma_u^2 = 3.37$ and $\sigma_\alpha^2 = 0.18$, and thus, we had $\sigma_x^2 = 3.55$ and $\rho_x = 0.05$. Then, to generate Y_{ijl}, we set $\sigma_b^2 = 0.20, \beta_0 = 0$ and $\beta_1 = 1.36$ and varied $\beta_{2(0)}$ and $\beta_{2(1)}$ across the four scenarios to obtain the value of success rates $\pi_0 = 0.50$ and $\pi_1 = 0.70$ in the control and intervention groups, respectively, on average over 1000 datasets. The value of ICC for outcome is expected to be different in the control and intervention groups because, for binary outcome, ICC depends on the success rate [26]. We used the expression $\rho_i = \text{Var}\left(\pi_{ij}\right) / \left(\pi_i(1-\pi_i)\right)$ [6, 27], where π_{ij} is the true proportion of success in the (ij)th cluster, to estimate the value of ICC for the ith intervention group. Firstly, we estimated Var $\left(\pi_{ij}\right)$ from a very big dataset with large number of clusters in each intervention group and with large cluster size. Then, with the success rates stated earlier for the control and intervention groups, the estimated ICC for outcome in the control and intervention groups were 0.037 and 0.032, respectively. We varied the number of clusters in each intervention group as $k = (5, 10, 20, 50)$ and fixed the cluster size $m = 50$. In the simulation studies, the four scenarios considered were **(S1)** $\beta_{2(0)} = \beta_{2(1)} = 1$ and $\psi_0 = \psi_1 = -1.34$; that is, both intervention groups have the same covariate effects on outcome and the same missingness mechanisms; **(S2)** $\beta_{2(0)} = \beta_{2(1)} = 1$ and $\psi_0 = -1.34, \psi_1 = 0.65$; that is, both intervention groups have the same covariate effects on outcome but different missingness mechanisms; **(S3)** $\beta_{2(0)} = 0.588, \beta_{2(1)} = 1$ and $\psi_0 = \psi_1 = -1.34$; that is, both intervention groups have different covariate effects on outcome but the same missingness mechanisms; and **(S4)** $\beta_{2(0)} = 0.588, \beta_{2(1)} = 1$ and $\psi_0 = -1.34, \psi_1 = 0.65$; that is, both intervention groups have different covariate effects on outcome and different missingness mechanisms. In **S1** and **S3**, there were 30% missing outcomes in each of the two intervention groups, while in **S2** and **S4**, there were 30% missing outcomes in the control group and 60% missing outcomes in the intervention group.

5.2. Data analysis

Each generated full and incomplete datasets were then analysed by CL_U, CL_A, RELR and GEE. Missing outcomes were handled using CRA and MMI. We included the interaction between intervention and baseline covariate into the analysis models RELR and GEE in the case of **S3** and **S4**. The R packages **lme4** and **geepack** were used to fit RELR and GEE, respectively. We used MMI, with a RELR imputation model, so that the imputation model was correctly specified. For **S3** and **S4**, an interaction between intervention and baseline covariate was included in the imputation model. The R package `jomo` [28] was used to multiply impute each generated incomplete dataset 15 times, although this package uses probit link between the mean of the outcome and the linear predictor. Both links give similar results as long as individual-level probabilities of success are not too small and not too large. The algorithm used by `jomo` [28] is essentially the same used by the REALCOM-IMPUTE software for MMI, details of which can be found in [29]. We used 100 burn-in iterations, which through preliminary investigations, we found to be sufficient for convergence to the posterior distribution of the parameters of our imputation model, and thinning rate 25 to reduce the autocorrelation between successive draws. When fitting the GEE models using the package **geepack** in R, we encountered convergence problems (maximum of three times out of 1000 simulation runs) in the case of **S2** and **S4**. In such situation, we fitted GEE assuming independent correlation structure.

5.3. Simulation results

Figure 1 represents the average estimates of RD and coverage rates of nominal 95% confidence intervals over 1000 simulation runs using CL_U and CL_A with CRA and MMI for each of the four scenarios. The corresponding numerical results using full data, CRA and MMI are available in Table F1 in Appendix F of the Supporting Information. The RD estimates using full data and using MMI followed by cluster-level analyses were unbiased for each of the four scenarios. However, CRA estimates were biased using both the CL_U and CL_A for each of the four scenarios. These results support our derived analytical results for RD estimates in Section 4.1. Under scenario **S3**, the CRA estimates of RD using both the CL_U and CL_A were coincidentally close to the true value of RD. In further simulations, where the parameter values were changed, the corresponding estimates of RD, using both the CL_U and CL_A, were found to be biased (see Table F2 in Appendix F in the Supporting Information). As expected, the average estimated SEs of CL_A are smaller than that of CL_U, using full data, CRA and MMI. This is because the CL_A removes the differences between the outcome values of the two intervention groups that can be attributed to differences in the baseline covariate. MMI with adjusted DF estimates gave overcoverage for nominal 95% confidence intervals for small number of clusters in each intervention group.

Figure 2 shows the average estimates of log(RR) and coverage rates for nominal 95% confidence intervals over 1000 simulation runs using CL_U and CL_A with CRA and MMI for the all four considered scenarios. The corresponding numerical results using full data, CRA and MMI are available in Table F3 in Appendix F of the Supporting Information. The full data estimates of log(RR) using CL_U and CL_A were very close to the true value. However, our analytical result showed that CL_A estimates of RR are biased if the data are generated from a logit link model. In this simulation, CL_A estimates were close to the true value because of the parameters' configuration. In a further simulation, where the parameters' values were changed, the estimates of log(RR) using CL_A were found to be biased (see Table F4 in Appendix F in the Supporting Information). The MMI followed by cluster-level analyses estimates of log(RR) were unbiased for all four considered scenarios. The CRA estimates were biased using both CL_U and CL_A for all four considered scenarios. These results support our derived analytical results for RR in Section 4.2. MMI with adjusted DF estimates resulted in the overcoverage of nominal 95% confidence intervals for small number of clusters in each intervention group.

Recall that RELR estimates cluster-specific (also known as conditional) intervention effect, while GEE estimates population-averaged (also known as marginal) intervention effect. In this study, the simulation data were generated using a RELR model (equation (2)), where we set $\beta_1 = 1.36$, which can be interpreted as conditional (on cluster random effects and baseline covariate X) log(OR) of developing the event of interest in the intervention group compared with the control group. The corresponding marginal value of β_1 will be smaller because the general effect of using a population-averaged model over cluster-specific model is to attenuate the regression coefficient [27]. Table I displays the average estimates of the log(OR), their average estimated SE and coverage rates of nominal 95% confidence intervals using RELR and GEE. The full data estimates of GEE is slightly lower as expected than that of RELR. For GEE, the CRA and MMI estimates were compared with the mean of the full data estimates as the true population-averaged

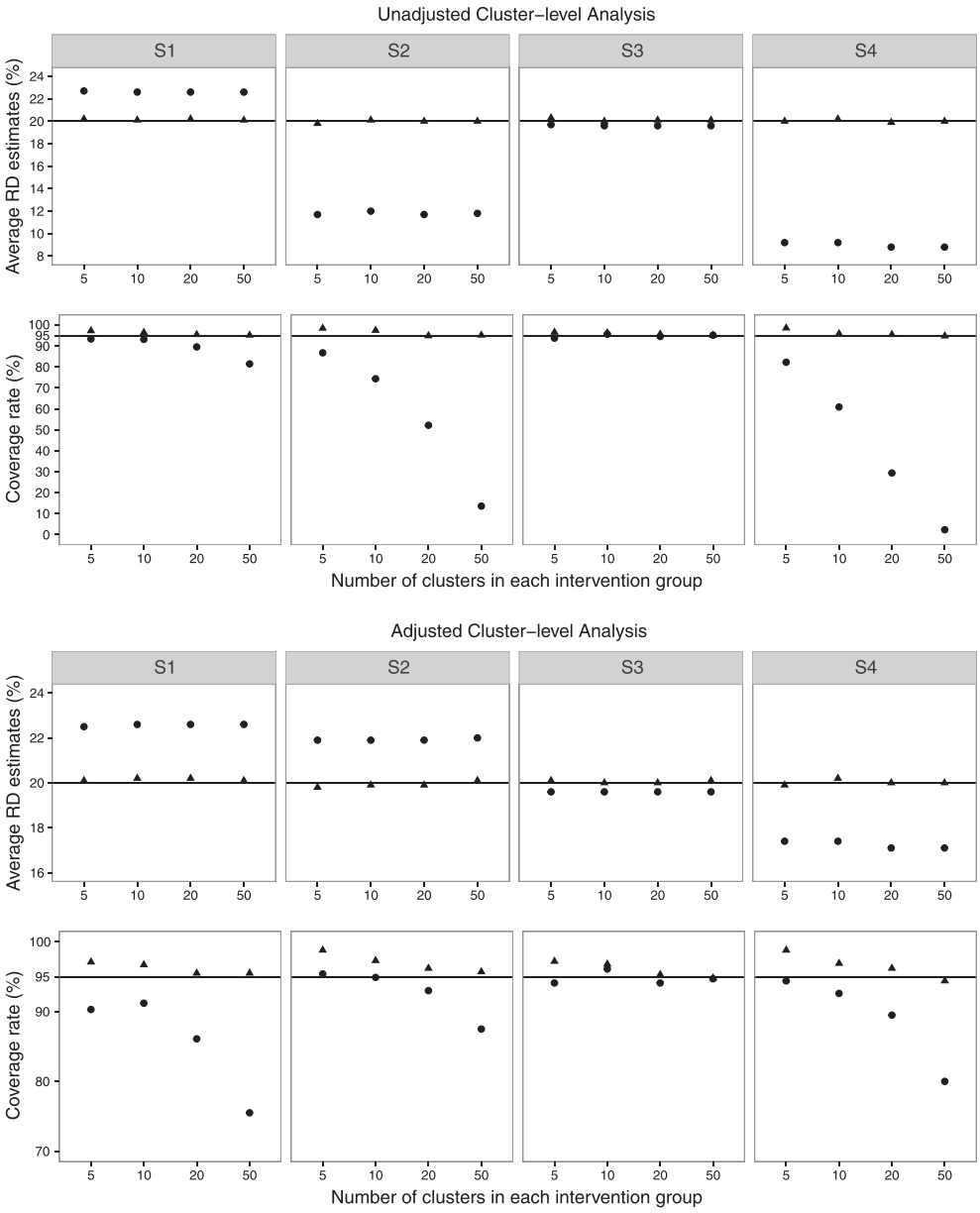

Figure 1. Simulation results for risk difference (RD). The columns represent the four scenarios considered in the simulation studies. The first and second rows represent the average estimates of RD and coverage rates for nominal 95% confidence interval, respectively, using unadjusted cluster-level analysis. The third and fourth rows represent the similar estimates using adjusted cluster-level analysis. Results are shown for complete records analysis (•) and multilevel multiple imputation (▲) over 1000 simulation runs. The lines (—) correspond to the true value.

log(OR) was unknown. The CRA estimates of RELR and GEE were unbiased with nominal coverage rates. This is because we were adjusting for the baseline covariate that was associated with missingness. However, RELR with MMI gave slightly upward biased (maximum 8.6%) estimates of intervention effect with small number of clusters in each intervention group, while GEE with MMI gave unbiased estimates. The study by Caille *et al.* [11] showed similar results to ours regarding good performance of GEE with respect to bias and coverage rate using MMI. The average estimated SEs of RELR estimates using CRA were lower than that of RELR using MMI, whereas the average estimated SEs of GEE estimates using CRA and MMI are fairly similar. Therefore, there is no benefit in doing MMI over CRA when the CRA and MMI use the same functional form of baseline covariates.

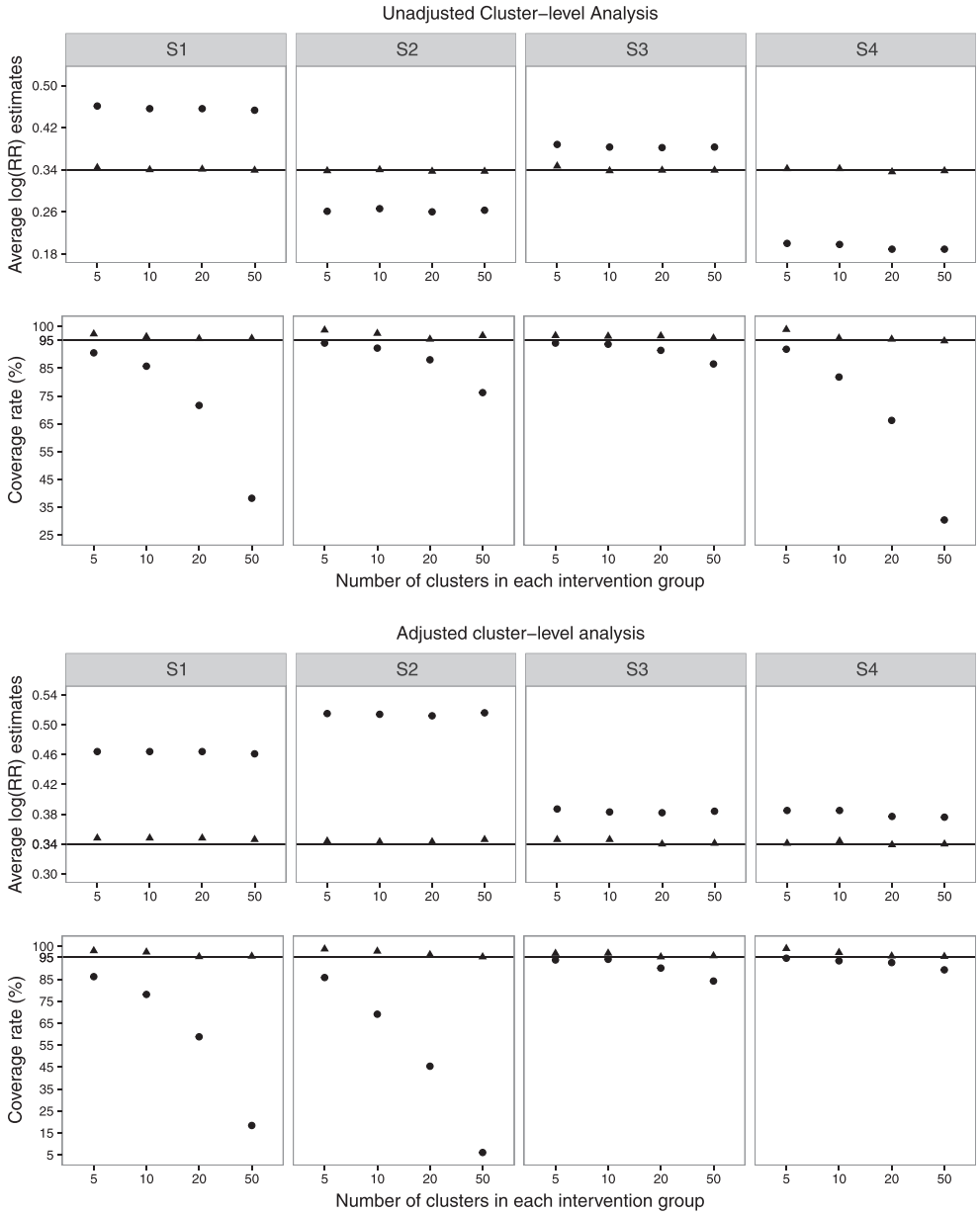

Figure 2. Simulation results for risk ratio (RR). The columns represent the four scenarios considered in the simulation studies. The first and second rows represent the average estimates of log(RR) and coverage rates for nominal 95% confidence interval, respectively, using unadjusted cluster-level analysis. The third and fourth rows represent the similar estimates using adjusted cluster-level analysis. Results are shown for complete records analysis (•) and multilevel multiple imputation (▲) over 1000 simulation runs. The lines (—) correspond to the true value.

6. Example

We now illustrate the methods compared here using the data from Health and Literacy Intervention (HALI) trial, a factorial CRT designed to investigate the impact of two interventions among school children in class 1 and class 5 on the south coast of Kenya [30]. The interventions were intermittent screening and treatment (IST) for malaria on the health and education of school children in class 1 and class 5 and a literacy intervention (LIT) on education only being applied in class 1. One hundred and one government primary schools were randomised to one of the four groups receiving (i) IST alone (25 schools); (ii) LIT alone (25 schools); (iii) both IST and LIT (26 schools); or (iv) neither IST nor LIT (25 schools). On average, the number of children per school in the four groups were, respectively, 107 (standard deviation (SD) = 7.54), 99 (SD = 17.84), 103 (SD = 6.28) and 102 (SD = 7.51). The primary outcomes were

anaemia at either 12 or 24 months and educational achievement at 9 and 24 months assessed by a battery of tests of reading, writing and arithmetic. Baseline characteristics of the school (school mean exam score and school size), the child (age, sex, sleep under net and baseline anaemia) and the household (paternal education and household size) were collected. For the purpose of illustration, we restricted attention to anaemia (binary) measured at the 24 months follow-up. A paper published based on this study [30] showed no evidence of interaction between the two interventions in class 1 where both were implemented. We therefore merged groups (i) and (iii) where IST was implemented and considered this as the intervention group and merged groups (ii) and (iv) where IST was not implemented and considered this as the control group. The control group and the intervention group consisted of 2502 and 2674 children, respectively, and among them, 475 (18.98%) and 501 (18.74%) had missing anaemia at 24 months, respectively. The covariate baseline anaemia had some missing values as well. To illustrate our methods for the case where only outcomes are missing and all baseline covariates are fully observed, we excluded the children from the analysis with missing baseline anaemia value. Hence, in our analysis, the control group and the intervention group consisted of 2373 and 2451 children, respectively, and among them, 430 (18.12%) and 424 (17.30%) had missing anaemia at 24 months, respectively.

The original trial's prespecified analysis planned to adjust for the baseline covariates' age, sex, exam score, literacy group and baseline anaemia. In our analysis, firstly, we investigated the association of the baseline covariates (age, sex, exam score, literacy group and baseline anaemia) with anaemia at 24 months and with the probability of anaemia outcome at 24 months being missing by fitting RELR models (see Table F5 in Appendix F of the Supporting Information). Age and baseline anaemia were strongly associated with anaemia at 24 months, and there was no evidence of interaction between IST intervention and baseline covariates in the model for anaemia at 24 months. Older children were more likely to have anaemia at 24 months missing, and children receiving LIT were less likely to have anaemia at 24 months missing. There was weak evidence of interaction between IST intervention and literacy group on the missingness of anaemia at 24 months. Based on these analyses, a working assumption is that missingness of anaemia at 24 months depends mainly on age and that this dependence does not differ between the two intervention groups as there was no evidence of interaction between IST intervention and age.

We analysed the data using the methods CL_U, CL_A, RELR and GEE, assuming that the missingness in anaemia at 24 months depends on the baseline covariates, but conditioning on these, not on the anaemia at 24 months itself, that is, a CDM mechanism. GEE models were fitted assuming both logit and log links for the true outcome model to estimate OR and RR, respectively. The objective of fitting GEE with log link was to estimate RR using individual-level analysis and to compare these estimates with the similar estimates obtained using cluster-level analyses. In addition, we wanted to compare our estimates of RR using GEE with the estimates of RR reported in the original paper [30] published based on this HALI trial data. The missing anaemia data at 24 months were handled using CRA and MMI. The RELR, GEE and adjusted cluster-level analyses were adjusted for the baseline covariates age, sex, school mean exam score, literacy group and baseline anaemia. MMI was carried out using the R package `jomo` [28], with an imputation model adjusted for the aforementioned baseline covariates. We used 100 imputed datasets in MMI. GEE with log link after MMI was not congenial with the imputation model, as the imputation model used probit link. The estimates and confidence intervals of RD, RR and OR obtained by CRA and MMI are displayed in Table II. Columns M_0 and M_1 in Table II represent the number of children in the control and intervention groups, respectively. All measures showed no evidence of IST intervention effect in improving health of school children by alleviating anaemia. The CRA estimates of RD and RR using cluster-level analyses are very similar to the corresponding estimates obtained by MMI. This is because CRA is valid in this case as there is no evidence of intervention effect and no evidence of interaction between covariates and intervention. The estimates and CIs of unadjusted and adjusted OR obtained by CRA were found to be very close to the corresponding estimates obtained by MMI. This is because, as we found in our simulation results, there is no gain in terms of bias or efficiency of the estimates using MMI over CRA as long as the same functional form of the same set of predictors of missingness are used by both methods.

7. Discussion and conclusion

In this paper, we showed analytically and through simulations that cluster-level analyses for estimating RD using complete records are valid only when there is no intervention effect in truth and the intervention groups have the same missingness mechanism and the same covariate effect in the outcome model. For estimating RR, cluster-level analyses using complete records are valid if the true data generating model

Table I. Average estimates of log(OR), their average estimated SEs and coverage rates for nominal 95% confidence intervals over 1000 simulation runs, using RELR and GEEs with full data, CRA and MMI. Monte Carlo errors for average estimates and average estimated SEs are all less than 0.016 and 0.003, respectively. The true value of conditional log(OR) in RELR is 1.36. The true value of population-averaged log(OR) for GEE was empirically estimated using full data.

	k	Average estimate						Average estimated SE						Coverage (%)					
		Full		CRA		MMI		Full		CRA		MMI		Full		CRA		MMI	
		RELR	GEE	RELR	GEE	RELR	GEE	RELR	GEE	RELR	GEE	RELR	GEE	RELR	GEE	RELR	GEE	RELR	GEE
S1	5	1.363	1.321	1.360	1.320	1.384	1.328	0.341	0.363	0.364	0.382	0.391	0.372	94.6	95.2	94.4	94.7	97.7	96.5
	10	1.365	1.321	1.368	1.323	1.392	1.329	0.252	0.258	0.268	0.271	0.284	0.272	94.6	95.2	94.4	95.1	96.1	96.0
	20	1.361	1.315	1.363	1.317	1.385	1.322	0.182	0.184	0.193	0.192	0.201	0.195	94.7	95.0	95.0	94.7	95.8	95.5
	50	1.359	1.310	1.361	1.310	1.380	1.316	0.118	0.117	0.125	0.122	0.129	0.124	94.4	95.1	94.8	95.4	94.8	95.0
S2	5	1.345	1.311	1.368	1.333	1.402	1.335	0.336	0.320	0.405	0.417	0.456	0.438	94.7	94.8	95.5	94.9	98.6	98.6
	10	1.350	1.309	1.356	1.313	1.384	1.308	0.250	0.258	0.298	0.301	0.330	0.317	93.2	94.4	94.7	95.4	97.0	97.1
	20	1.358	1.311	1.352	1.305	1.376	1.301	0.184	0.185	0.215	0.213	0.232	0.224	94.8	95.8	95.0	94.9	96.7	96.4
	50	1.366	1.316	1.367	1.318	1.389	1.316	0.118	0.117	0.138	0.135	0.146	0.141	95.3	95.7	95.0	95.0	95.8	96.0
S3	5	1.391	1.353	1.407	1.367	1.434	1.374	0.343	0.358	0.392	0.400	0.414	0.389	94.8	94.1	95.2	94.4	97.7	97.4
	10	1.352	1.307	1.359	1.314	1.385	1.320	0.254	0.259	0.284	0.286	0.299	0.285	92.8	94.1	94.0	94.5	95.4	95.0
	20	1.372	1.326	1.370	1.325	1.395	1.330	0.183	0.184	0.204	0.202	0.212	0.203	93.2	94.4	93.2	94.1	94.1	94.1
	50	1.363	1.313	1.363	1.313	1.386	1.317	0.118	0.117	0.132	0.127	0.135	0.129	95.1	95.1	94.8	95.5	95.4	95.4
S4	5	1.375	1.336	1.413	1.378	1.476	1.390	0.346	0.366	0.497	0.493	0.535	0.505	94.5	95.2	97.0	94.0	98.6	98.5
	10	1.366	1.325	1.377	1.334	1.431	1.342	0.252	0.258	0.353	0.351	0.375	0.357	94.6	95.3	95.3	94.6	96.5	96.6
	20	1.376	1.328	1.387	1.339	1.432	1.346	0.183	0.184	0.252	0.247	0.266	0.251	94.7	94.8	94.3	94.4	94.5	94.8
	50	1.360	1.312	1.362	1.313	1.397	1.317	0.118	0.117	0.160	0.156	0.167	0.157	95.4	95.7	94.8	94.5	94.4	94.2

SEs: standard errors; RELR: random effects logistic regression; GEE: generalised estimation equations; CRA: complete records analysis; MMI: multilevel multiple imputation.

Table II. Risk difference, risk ratio and odds ratio estimates using CRA and MMI for the IST intervention trial data.

Analysis approach	M_0	M_1	Risk difference	Risk ratio	Odds ratio
			Estimate (95% CI)	Estimate (95% CI)	Estimate (95% CI)
Cluster-level analysis[a]					
CRA					
Unadjusted	2027	2173	0.019 (−0.040, 0.077)	1.047 (0.908, 1.208)	
Adjusted	1935	2027	0.022 (−0.033, 0.077)	1.037 (0.908, 1.185)	
MMI					
Unadjusted	2373	2451	0.021 (−0.038, 0.080)	1.053 (0.911, 1.218)	
Adjusted	2373	2451	0.017 (−0.035, 0.070)	1.040 (0.910, 1.189)	
Individual-level analysis					
CRA					
RELR					
Unadjusted	2027	2173		—	1.090 (0.841, 1.414)
Adjusted	1935	2027		—	1.088 (0.839, 1.409)
GEE[b]					
Unadjusted	2027	2173		1.048 (0.908, 1.209)	1.082 (0.850, 1.378)
Adjusted	1935	2027		1.019 (0.911, 1.141)	1.070 (0.842, 1.359)
MMI					
RELR					
Unadjusted	2373	2451		—	1.101 (0.849, 1.428)
Adjusted	2373	2451		—	1.089 (0.841, 1.413)
GEE					
Unadjusted	2373	2451		1.053 (0.912, 1.215)	1.090 (0.856, 1.389)
Adjusted	2373	2451		1.019 (0.911, 1.140)	1.072 (0.843, 1.363)

[a]Cluster-level analysis was used to estimate the risk difference and the risk ratio.
[b]GEE was used to estimate the risk ratio using log link and to estimate the marginal odds ratio using logit link.
CRA, complete records analysis; MMI, multilevel multiple imputation; RELR, random effects logistic regression; GEE, generalised estimation equation; IST, intermittent screening and treatment; CI, confidence interval.

has log link and the intervention groups have the same missingness mechanism and the same covariate effect in the outcome model. However, if the true data generating model has logit link, cluster-level analyses using complete records for estimating RR are valid only when there is no intervention effect in truth and the intervention groups have the same missingness mechanism and the same covariate effect in the outcome model. But, in practice, it is impossible to know in advance whether there is an intervention effect. We therefore caution researchers that cluster-level analyses using complete records, assuming logit link for the true data generating model, in general results in biased inferences for RR in CRTs. However, when the true data generating model follows a log link and the parameter of interest is RR, cluster-level analyses using complete records give valid inferences if the intervention groups have the same missingness mechanism and the same covariates effect in the outcome model.

In contrast, MMI followed by cluster-level analyses gave unbiased estimates of RD and RR regardless of whether missingness mechanisms were the same or different between the intervention groups and whether there is an interaction between intervention and baseline covariate in the outcome model, provided that an interaction was allowed for in the imputation model when required. However, MMI resulted in overcoverage for the nominal 95% confidence interval with small number of clusters in each intervention group. Similar results were found for continuous outcomes in CRTs by Hossain *et al.* [13].

The full data estimates of conditional (on cluster random effects and covariates) log(OR) using RELR were unbiased with good coverage rates. These results differ from the results found by Ma *et al.* [10], where they concluded that full data estimates using RELR were biased. As noted previously, we believe this is because they generated the data in such a way that they knew what the true population-averaged log(OR) was, but after fitting RELR, they compared the estimates of conditional log(OR) with the true

population-averaged log(OR). As noted earlier, population-averaged log(OR) is marginal with respect to the cluster random effects [31].

The CRA estimates of conditional log(OR) using RELR were unbiased with coverage rates close to the nominal level regardless of whether the missingness mechanism is the same or different between the intervention groups and whether there is an interaction between the intervention and baseline covariate in the data generating model for outcome, provided that if there is an interaction in the data generating model for the outcome, then this interaction is included in the model fitted to the data. This conclusion contradicts the results of a previous study by Ma *et al.* [10], where they found that CRA estimates using RELR are biased under CDM assumption. Again we believe this is because they compared RELR estimates of the conditional log(OR) with the true marginal log(OR). The conclusions of Ma *et al.* [10] have subsequently been cited in a recent textbook on CRT design and analysis [27]. We hope that our results and explanations help in understanding some of the surprising results and conclusion in Ma *et al.* [8–10]. In our study, we also found that the RELR with MMI gave slightly upward biased estimates of conditional log(OR) for small number of clusters in each intervention groups.

The GEE using CRA and MMI gave unbiased estimates of population-averaged log(OR) with coverage rates close to the nominal level regardless of whether the missingness mechanism was the same between the intervention groups and whether there was an interaction between the intervention group and baseline covariate in the data generating model. Similar results had been found by Ma *et al.* [10] for GEE in terms of bias, although as described earlier, in their data generating mechanism, the covariate was generated independently of the outcome.

In this study, we assumed the same covariate effects for the probability of having a missing outcome in the two intervention groups. Another possible scenario would be that the two groups have different missingness mechanism in the sense that the covariate effects on the probability of having missing outcome are different between the two intervention groups. To address this, we have carried out a further simulation with different covariate effects ($\phi_0 = 0.5, \phi_1 = 1$) on the probability of having a missing outcome between the two groups. The results showed, as expected by theory, that CRA gives valid estimates. This is because, CRA is valid as long as conditional on the covariates in the model, the missingness is independent of the outcome. We also assumed baseline CDM assumption for binary outcome, which is an example of MAR as our baseline covariate was fully observed. In practice, it cannot be identified on the basis of the observed data that missingness assumption is appropriate [32, 33]. Therefore, sensitivity analyses should be performed [33, Ch. 10] to explore whether inferences are robust to the primary working assumption regarding the missingness mechanism. Furthermore, we focused on studies with only one individual-level baseline covariate; the methods described can be extended to more than one baseline covariate.

In conclusion, as long as both MMI and CRA use the same covariates with the same functional form, RELR or GEE using complete records can be recommended as the primary analysis approach for CRTs with missing binary outcomes if we are willing to assume that the missingness depends on baseline covariates and conditional on these, not on the outcome. In addition, where the aim is to estimate RD or RR, MMI can be used followed by cluster-level analysis to acquire valid estimates under the CDM assumption for missing binary outcomes, but one should be cautious when making inferences as this approach results in overcoverage for small number of clusters in each intervention group.

Acknowledgements

A. Hossain was supported by the Economic and Social Research Council (ESRC), UK, via Bloomsbury Doctoral Training Centre (ES/J5000021/1). K. DiazOrdaz was funded by the Medical Research Council (MRC) career development award in Biostatistics (MR/L011964/1). J. W. Bartlett's contribution to this paper was partly supported by MRC fellowship (MR/K02180X/1) while he was a member of the Department of Medical Statistics, London School of Hygiene and Tropical Medicine (LSHTM).

The dataset used as example in this paper was made available through the HALI trial, with study design and data collection by the team in Kenya led by Simon Brooker, Katherine Halliday, Matthew Jukes, George Okello and Carlos Mcharo, and funded by 3ie, the Partnership for Child Development, and the Development Impact Evaluation Initiative as part of the Malaria Impact Evaluation Program of the World Bank. The authors would like to thank Professor Elizabeth Allen, Department of Medical Statistics, LSHTM, for helping us obtaining permission to use HALI trial data. We also would like to thank all the teachers, children and parents who participated in this trial.

References

1. Campbell M, Donner A, Klar N. Developments in cluster randomized trials and statistics in medicine. *Statistics in medicine* 2007; **26**(1):2–19.
2. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold: London, 2000.
3. Murray DM, Blitstein JL. Methods to reduce the impact of interclass correlation in group-randomised trials. *Evaluation Review* 2003; **27**:79–103.
4. Murray DM. *Design and Analysis of Group-randomized Trials*. Oxford University Press: New York, 1998.
5. Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. *American Journal of Public Health* 2004; **94**(3):423–432.
6. Hayes RJ, Moulton LH. *Cluster Randomised Trials*. CRC Press, Taylor & Francis Group, 2009.
7. Diaz-Ordaz K, Kenward MG, Cohen A, Coleman CL, Eldridge S. Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. *Clinical Trials* 2014; **11**(5):590–600.
8. Ma J, Akhtar-Danesh N, Dolovich L, Thabane L, the CHAT investigators. Imputation strategies for missing binary outcomes in cluster randomized trials. *BMC Medical Research Methodology* 2011; **11**:18.
9. Ma J, Raina P, Beyene J, Thabane L. Comparing the performance of different multiple imputation strategies for missing binary outcomes in cluster randomized trials: a simulation study. *J Open Access Med Stat* 2012; **2**:93–103.
10. Ma J, Raina P, Beyene J, Thabane L. Comparison of population-averaged and cluster-specific models for the analysis of cluster randomized trials with missing binary outcomes: a simulation study. *BMC Medical Research Methodology* 2013; **13**:9.
11. Caille A, Leyrat C, Giraudeau B. A comparison of imputation strategies in cluster randomized trials with missing binary outcomes. *Statistical Methods in Medical Research* 2016; **25**(6):2650–2669.
12. Groenwold RH, Donders AR, Roes KC, Harrell FE, Moons KG. Dealing with missing outcome data in randomized trials and observational studies. *American Journal of Epidemiology* 2012; **175**(3):210–217.
13. Hossain A, Diaz-Ordaz K, Bartlett JW. Missing continuous outcomes under covariate dependent missingness in cluster randomised trials. *Statistical Methods in Medical Research* 2016. https://doi.org/10.1177/0962280216648357.
14. Davies HTO, Crombie IK, Tavakoli M. When can odds ratios mislead *BMJ* 1998; **316**(7136):989–991.
15. Hernandez AV, Steyerberg EW, Habbema JD. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *Journal of Clinical Epidemiology* 2004; **57**(5): 454–460.
16. Gail M, Tan W, Piantadosi S. Tests for no treatment effect in randomised clinical trials. *Biometrika* 1988; **75**(1):57–64.
17. Li P, Redden DT. Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analyzing binary outcome in small sample cluster-randomized trials. *BMC Medical Research methodology* 2015; **15**(1):38.
18. Ukoumunne OC, Carlin JB, Gulliford MC. A simulation study of odds ratio estimation for binary outcomes from cluster randomized trials. *Statistics in Medicine* 2007; **26**(18):3415–3428.
19. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. *Biometrics* 2001; **57**(1):126–134.
20. Little RJA, Rubin DB. *Statistical Analysis with Missing Data* 2nd ed. John Wiley & Sons: New Jersey, 2002.
21. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons: New York, 1987.
22. Meng XL. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* 1994; **9**(4):538–558.
23. Barnard J, Rubin DB. Small-sample degrees of freedom with multiple imputation. *Biometrika* 1999; **86**(4):948–955.
24. Andridge RR. Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical Journal* 2011; **53**(1):57–74.
25. DiazOrdaz K, Kenward M, Gomes M, Grieve R. Multiple imputation methods for bivariate outcomes in cluster randomised trials. *Statistics in Medicine* 2016; **35**(20):3482–3496.
26. Gulliford M, Adams G, Ukoumunne O, Latinovic R, Chinn S, Campbell M. Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. *Journal of clinical epidemiology* 2005; **58**(3):246–251.
27. Campbell MJ, Walters SJ. *How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research*. John Wiley & Sons, 2014.
28. Quartagno M, Carpenter J. *jomo: a package for multilevel joint modelling multiple imputation*, 2015. http://CRAN.R-project.org/package=jomo.
29. Carpenter JR, Goldstein H, Kenward MG. REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software* 2011; **45**(5):1–14.
30. Halliday KE, Okello G, Turner EL, Njagi K, Mcharo C, Kengo J, Allen E, Dubeck MM, Jukes MC, Brooker SJ. Impact of intermittent screening and treatment for malaria among school children in Kenya: a cluster randomised trial. *PLoS Med* 2014; **11**(1):e1001594.
31. Faraway J. *Extending the Linear Model with R*. Taylor & Francis Group, 2006.
32. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine* 2010; **29**(28):2920–2931.
33. Carpenter JR, Kenward MG. *Multiple Imputations and Its Applications*. John Wiley & Sons, 2013.

Supporting information

Additional supporting information may be found online in the supporting information tab for this article.

Missing binary outcomes under covariate dependent missingness in cluster randomised trials

Anower Hossain[1, 2], Karla DiazOrdaz[1], and Jonathan W. Bartlett [3]

[1]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK.
[2]Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka, Bangladesh.
[3]Statistical Innovation Group, AstraZeneca.

Appendix A

In this appendix, we show that, with full data, $\widehat{\text{RD}}_{\text{unadj}}$ is unbiased for true RD, and $\widehat{\text{RR}}_{\text{unadj}}$ is consistent (and, therefore, asymptotically unbiased) for true RR. We have

$$\text{E}\left(\bar{p}_i\right) = \text{E}\left(\frac{1}{k}\sum_{i=1}^{k} p_{ij}\right) = \frac{1}{mk}\sum_{j=1}^{k}\sum_{l=1}^{m}\text{E}\left(Y_{ijl}\right) = \pi_i$$

where π_i is the true proportion of success in the ith intervention group. Then

$$\text{E}\left(\widehat{\text{RD}}_{\text{unadj}}\right) = \text{E}\left(\bar{p}_1 - \bar{p}_0\right) = \pi_1 - \pi_0 = \text{RD}.$$

Hence $\widehat{\text{RD}}_{\text{unadj}}$ is unbiased for true RD.

Now, since

$$\bar{p}_0 \xrightarrow{prob.} \pi_0 \quad \text{and} \quad \bar{p}_1 \xrightarrow{prob.} \pi_1 \quad \text{as} \quad k \to \infty,$$

$$\widehat{\text{RR}}_{\text{unadj}} = \frac{\bar{p}_1}{\bar{p}_0} \xrightarrow{prob.} \frac{\pi_1}{\pi_0} = \text{RR} \quad \text{as} \quad k \to \infty.$$

Therefore, $\widehat{\text{RR}}_{\text{unadj}}$ is consistent (and, therefore, asymptotically unbiased) for true RR as $k \to \infty$.

Appendix B

In this appendix, we show that the adjusted cluster-level estimator of risk ratio (RR) with full data is a consistent estimator (and, therefore, asymptotically unbiased) of true RR under certain conditions.

As we defined in equation (5) in the main paper, the adjusted cluster-level estimator of RR is given by

$$\widehat{\mathrm{RR}}_{\mathrm{adj}} = \frac{\bar{\epsilon}_1^r}{\bar{\epsilon}_0^r} = \frac{\frac{1}{k}\sum_{j=1}^{k} \frac{N_{1j}}{\hat{N}_{1j}}}{\frac{1}{k}\sum_{j=1}^{k} \frac{N_{0j}}{\hat{N}_{0j}}} \tag{B1}$$

If $k \to \infty$, the numerator is a consistent estimator of

$$\begin{aligned}
\mathrm{E}\left(\frac{N_{1j}}{\hat{N}_{1j}}\right) &= \mathrm{E}\left[\mathrm{E}\left(\frac{N_{1j}}{\hat{N}_{1j}}\middle|\delta_{1j}, \boldsymbol{X}_{1j}\right)\right] \\
&= \mathrm{E}\left[\frac{\mathrm{E}\left(N_{1j}|\delta_{1j}, \boldsymbol{X}_{1j}\right)}{\hat{N}_{1j}(\boldsymbol{X}_{1j})}\right] \\
&= \mathrm{E}\left[\frac{\sum_{l=1}^{m} \pi_{1jl}}{\hat{N}_{1j}(\boldsymbol{X}_{1j})}\right]
\end{aligned}$$

Assuming the data are generated from the log link model given in equation (1) in the main paper, we have

$$\begin{aligned}
\mathrm{E}\left(\frac{N_{1j}}{\hat{N}_{1j}}\right) &= \mathrm{E}\left[\frac{\sum_{l=1}^{m} \exp\left(\beta_0 + \beta_1 + f_1(X_{1jl}) + \delta_{1j}\right)}{\hat{N}_{1j}(\boldsymbol{X}_{1j})}\right] \\
&= \exp\left(\beta_0 + \beta_1\right) \mathrm{E}\left[\frac{\exp\left(\delta_{1j}\right) \sum_{l=1}^{m} \exp\left(f_1(X_{1jl})\right)}{\hat{N}_{1j}(\boldsymbol{X}_{1j})}\right]
\end{aligned} \tag{B2}$$

Similarly, it can be shown that, if $k \to \infty$, the denominator of equation (B1) is a consistent estimator of

$$\begin{aligned}
\mathrm{E}\left(\frac{N_{0j}}{\hat{N}_{0j}}\right) &= \mathrm{E}\left[\frac{\sum_{l=1}^{m} \exp\left(\beta_0 + f_0(X_{0jl}) + \delta_{0j}\right)}{\hat{N}_{0j}(\boldsymbol{X}_{0j})}\right] \\
&= \exp\left(\beta_0\right) \mathrm{E}\left[\frac{\exp\left(\delta_{0j}\right) \sum_{l=1}^{m} \exp\left(f_0(X_{0jl})\right)}{\hat{N}_{0j}(\boldsymbol{X}_{0j})}\right]
\end{aligned} \tag{B3}$$

The distribution of X (in expectation) is the same between the intervention groups as a consequence of randomisation. If δ_{0j} and δ_{1j} have common distribution, and $f_i(X_{ijl}) = f(X_{ijl})$ for $i \in \{0, 1\}$, the expectations in the right hand side of equations (B2) and (B3) are equal. Hence, we have

$$\widehat{\mathrm{RR}}_{\mathrm{adj}} \to \exp(\beta_1) = \mathrm{RR} \qquad \text{as} \qquad k \to \infty.$$

Therefore, the adjusted cluster-level estimator of RR is consistent and, therefore, asymptotically unbiased (as $k \to \infty$) for true RR if (i) the true data generating model is a log link model, (ii) the functional form of the covariates is the same between the intervention groups, and (iii) the distribution of random effect is the same between the intervention groups.

The above argument is not true if the data are generating from the logit link model (2) in the main paper with $\beta_1 \neq 0$, and, therefore, $\widehat{\mathrm{RR}}_{\mathrm{adj}}$ is not consistent for true RR ($\neq 1$). However, under the null hypothesis of no intervention effect ($\beta_1 = 0$), the above argument is true if the true data generating model has logit link. Hence $\widehat{\mathrm{RR}}_{\mathrm{adj}}$ is consistent for true RR ($= 1$) as $k \to \infty$.

Appendix C

In this appendix we show that the cluster-level analyses for RD using CRA are biased. To this end, we write the individual-level probabilities of success, π_{ijl}, as

$$\pi_{ijl} = \pi_i + g_i\left(X_{ijl}, \delta_{ij}\right)$$

where $g_i\left(X_{ijl}, \delta_{ij}\right)$ is a function of baseline covariate X_{ijl} and random cluster-effect δ_{ij}, and which determines how individual-level probabilities of success differ from group level probability of success in each intervention group. Then

$$\mathrm{E}_{j,l}\left(\pi_{ijl}|R_{ijl}=1\right) = \pi_i + \mathrm{E}_{j,l}\left(g_i\left(X_{ijl}, \delta_{ij}\right)|R_{ijl}=1\right)$$

and

$$\begin{aligned}
\mathrm{E}\left(\widehat{\mathrm{RD}}^{\mathrm{cr}}_{\mathrm{unadj}}\right) &= \mathrm{E}\left(\pi_{1jl}|R_{1jl}=1\right) - \mathrm{E}\left(\pi_{0jl}|R_{0jl}=1\right) \\
&= \pi_1 - \pi_0 + \mathrm{E}\left(g_1\left(X_{1jl}, \delta_{1j}\right)|R_{1jl}=1\right) - \mathrm{E}\left(g_0\left(X_{0jl}, \delta_{0j}\right)|R_{0jl}=1\right) \\
&= \mathrm{RD} + \mathrm{E}\left(g_1\left(X_{1jl}, \delta_{1j}\right)|R_{1jl}=1\right) - \mathrm{E}\left(g_0\left(X_{0jl}, \delta_{0j}\right)|R_{0jl}=1\right).
\end{aligned}$$

So $\widehat{\mathrm{RD}}^{\mathrm{cr}}_{\mathrm{unadj}}$ will be unbiased for true RD if and only if

$$\mathrm{E}\left(g_1\left(X_{1jl}, \delta_{1j}\right)|R_{1jl}=1\right) = \mathrm{E}\left(g_0\left(X_{0jl}, \delta_{0j}\right)|R_{0jl}=1\right).$$

Assuming the data are generated from the log link model (1) in the main paper, we have

$$g_i\left(X_{ijl}, \delta_{ij}\right) = \pi_{ijl} - \pi_i = \exp(\beta_0 + \beta_1\, i)\left\{\exp\left(f_i(X_{ijl}) + \delta_{ij}\right) - \mathrm{E}_{j,l}\left(\exp\left(f_i(X_{ijl}) + \delta_{ij}\right)\right)\right\} \tag{C1}$$

since $\pi_i = \mathrm{E}_{j,l}\left(\pi_{ijl}\right)$. If there is an intervention effect in truth ($\beta_1 \neq 0$), in general, we have from (C1)

$$\mathrm{E}\left(g_1\left(X_{1jl}, \delta_{1j}\right)|R_{1jl}=1\right) \neq \mathrm{E}\left(g_0\left(X_{0jl}, \delta_{0j}\right)|R_{0jl}=1\right)$$

even if the two intervention groups have the same missingness mechanism and the same covariate effects in the data generating model for the outcome. Hence, $\widehat{\mathrm{RD}}^{\mathrm{cr}}_{\mathrm{unadj}}$ is biased for true RD when the true data generating model has log link. However, under the null hypothesis of no intervention effect ($\beta_1 = 0$), if the two intervention groups have the same covariate effect, i.e. $f_i(X_{ijl}) = f(X_{ijl})$ for $i \in \{0, 1\}$, we have

$$g_i\left(X_{ijl}, \delta_{ij}\right) = \exp(\beta_0)\left\{\exp\left(f(X_{ijl}) + \delta_{ij}\right) - \mathrm{E}_{j,l}\left(\exp\left(f(X_{ijl}) + \delta_{ij}\right)\right)\right\}$$

and then, in addition, if the two intervention groups have the same missingness mechanism, we have

$$\mathrm{E}\left(g_1\left(X_{1jl}, \delta_{1j}\right)|R_{1jl}=1\right) = \mathrm{E}\left(g_0\left(X_{0jl}, \delta_{0j}\right)|R_{0jl}=1\right)$$

and hence $\widehat{\mathrm{RD}}^{\mathrm{cr}}_{\mathrm{unadj}}$ is unbiased for true RD = 0.

On the other hand, if we assume the data are generated from the logit link model (2) in the main paper, we have

$$\begin{aligned} g_i\left(X_{ijl}, \delta_{ij}\right) &= \pi_{ijl} - \pi_i \\ &= \operatorname{expit}\left(\beta_0 + \beta_1 i + f_i(X_{ijl}) + \delta_{ij}\right) - \mathrm{E}_{j,l}\left(\operatorname{expit}\left(\beta_0 + \beta_1 i + f_i(X_{ijl}) + \delta_{ij}\right)\right) \end{aligned} \tag{C2}$$

Then, again with $\beta_1 \neq 0$, we have from (C2)

$$\mathrm{E}\left(g_1\left(X_{1jl}, \delta_{1j}\right) | R_{1jl} = 1\right) \neq \mathrm{E}\left(g_0\left(X_{0jl}, \delta_{0j}\right) | R_{0jl} = 1\right)$$

even if the two intervention groups have the same missingness mechanism and the same covariate effect. Hence, $\widehat{\mathrm{RD}}^{\mathrm{cr}}_{\mathrm{unadj}}$ is biased for true RD when the true data generating model has logit link. However, like log link, under the null hypothesis of no intervention effect ($\beta_1 = 0$), if the two intervention groups have the same covariate effect, i.e. $f_i(X_{ijl}) = f(X_{ijl})$ for $i \in \{0, 1\}$ and if δ_{0j} and δ_{1j} have common distribution, we have

$$g_i\left(X_{ijl}, \delta_{ij}\right) = \operatorname{expit}\left(\beta_0 + f(X_{ijl}) + \delta_{ij}\right) - \mathrm{E}_{j,l}\left(\operatorname{expit}\left(\beta_0 + f(X_{ijl}) + \delta_{ij}\right)\right)$$

and then, in addition, if the two intervention groups have the same missingness mechanism, we have

$$\mathrm{E}\left(g_1\left(X_{1jl}, \delta_{1j}\right) | R_{1jl} = 1\right) = \mathrm{E}\left(g_0\left(X_{0jl}, \delta_{0j}\right) | R_{0jl} = 1\right)$$

and hence $\widehat{\mathrm{RD}}^{\mathrm{cr}}_{\mathrm{unadj}}$ is unbiased for true RD = 0.

Appendix D

In this appendix we investigate the validity of the cluster-level analyses for RR using CRA. To this end, we write π_{ijl} as

$$\pi_{ijl} = \pi_i \, h_i\left(X_{ijl}, \delta_{ij}\right)$$

where $h_i\left(X_{ijl}, \delta_{ij}\right)$ is a function of baseline covariate X_{ijl} and random cluster-effect δ_{ij}, and which determines how individual-level probabilities of success differ from group level probability of success. Then

$$\mathrm{E}_{j,l}\left(\pi_{ijl} | R_{ijl} = 1\right) = \pi_i \, \mathrm{E}_{j,l}\left(h_i\left(X_{ijl}, \delta_{ij}\right) | R_{ijl} = 1\right)$$

and

$$\begin{aligned} \widehat{\mathrm{RR}}^{\mathrm{cr}}_{\mathrm{unadj}} &\longrightarrow \frac{\mathrm{E}\left(\pi_{1jl} | R_{1jl} = 1\right)}{\mathrm{E}\left(\pi_{0jl} | R_{0jl} = 1\right)} \text{ as } k \longrightarrow \infty \\ &= \frac{\pi_1 \, \mathrm{E}\left(h_1\left(X_{1jl}, \delta_{1j}\right) | R_{1jl} = 1\right)}{\pi_0 \, \mathrm{E}\left(h_0\left(X_{0jl}, \delta_{0j}\right) | R_{0jl} = 1\right)} \\ &= \mathrm{RR} \, \frac{\mathrm{E}\left(h_1\left(X_{1jl}, \delta_{1j}\right) | R_{1jl} = 1\right)}{\mathrm{E}\left(h_0\left(X_{0jl}, \delta_{0j}\right) | R_{0jl} = 1\right)} \end{aligned}$$

So $\widehat{\mathrm{RR}}^{\mathrm{cr}}_{\mathrm{unadj}}$ will be consistent for true RR if only if

$$\frac{\mathrm{E}\left(h_1\left(X_{1jl},\delta_{1j}\right)|R_{1jl}=1\right)}{\mathrm{E}\left(h_0\left(X_{0jl},\delta_{0j}\right)|R_{0jl}=1\right)}=1.$$

Assuming the data are generated from the log link model (1) in the main paper, we have

$$\begin{aligned} h_i\left(X_{ijl},\delta_{ij}\right) &= \frac{\exp\left(\beta_0+\beta_1 i+f_i(X_{ijl})+\delta_{ij}\right)}{\mathrm{E}_{j,l}\left(\exp\left(\beta_0+\beta_1 i+f_i(X_{ijl})+\delta_{ij}\right)\right)} \\ &= \frac{\exp\left(f_i(X_{ijl})+\delta_{ij}\right)}{\mathrm{E}_{j,l}\left(\exp\left(f_i(X_{ijl})+\delta_{ij}\right)\right)} \end{aligned}$$

and

$$\frac{\mathrm{E}\left(h_1\left(X_{1jl},\delta_{1j}\right)|R_{1jl}=1\right)}{\mathrm{E}\left(h_0\left(X_{0jl},\delta_{0j}\right)|R_{0jl}=1\right)}=\frac{\mathrm{E}\left(\exp\left(f_1(X_{1jl})+\delta_{1j}\right)|R_{1jl}=1\right)}{\mathrm{E}\left(\exp\left(f_0(X_{0jl})+\delta_{0j}\right)|R_{0jl}=1\right)}\times\frac{\mathrm{E}\left(\exp\left(f_0(X_{0jl})+\delta_{0j}\right)\right)}{\mathrm{E}\left(\exp\left(f_1(X_{1jl})+\delta_{1j}\right)\right)}$$

Then if the two intervention groups have the same covariate effect, i.e. $f_i(X_{ijl})=f(X_{ijl})$ for $i\in\{0,1\}$) and if δ_{0j} and δ_{1j} have common distribution, we have

$$\frac{\mathrm{E}\left(\exp\left(f_0(X_{0jl})+\delta_{0j}\right)\right)}{\mathrm{E}\left(\exp\left(f_1(X_{1jl})+\delta_{1j}\right)\right)}=1$$

and, in addition, if the two intervention groups have the same missingness mechanism, we have

$$\frac{\mathrm{E}\left(\exp\left(f_1(X_{1jl})+\delta_{1j}\right)|R_{1jl}=1\right)}{\mathrm{E}\left(\exp\left(f_0(X_{0jl})+\delta_{0j}\right)|R_{0jl}=1\right)}=1$$

Therefore, if the two intervention groups have the same missingness mechanism and the same covariate effects, we have

$$\frac{\mathrm{E}\left(h_1\left(X_{1jl},\delta_{1j}\right)|R_{1jl}=1\right)}{\mathrm{E}\left(h_0\left(X_{0jl},\delta_{0j}\right)|R_{0jl}=1\right)}=1$$

and hence $\widehat{\mathrm{RR}}^{\mathrm{cr}}_{\mathrm{unadj}}$ is consistent for true RR.

On the other hand, assuming the data are generated from the logit link model (2) in the main paper, we have

$$h_i\left(X_{ijl},\delta_{ij}\right) = \frac{\mathrm{expit}\left(\beta_0+\beta_1 i+f_i(X_{ijl})+\delta_{ij}\right)}{\mathrm{E}_{j,l}\left(\mathrm{expit}\left(\beta_0+\beta_1 i+f_i(X_{ijl})+\delta_{ij}\right)\right)}$$

and

$$\begin{aligned} \frac{\mathrm{E}\left(h_1\left(X_{1jl},\delta_{1j}\right)|R_{1jl}=1\right)}{\mathrm{E}\left(h_0\left(X_{0jl},\delta_{0j}\right)|R_{0jl}=1\right)} &= \frac{\mathrm{E}\left(\mathrm{expit}\left(\beta_0+\beta_1+f_1(X_{1jl})+\delta_{1j}\right)|R_{1jl}=1\right)}{\mathrm{E}\left(\mathrm{expit}\left(\beta_0+f_0(X_{0jl})+\delta_{0j}\right)|R_{0jl}=1\right)} \\ &\quad\times\frac{\mathrm{E}\left(\mathrm{expit}\left(\beta_0+f_0(X_{0jl})+\delta_{0j}\right)\right)}{\mathrm{E}\left(\mathrm{expit}\left(\beta_0+\beta_1+f_1(X_{1jl})+\delta_{1j}\right)\right)} \end{aligned} \quad \text{(D1)}$$

If $\beta_1\neq 0$, we have

$$\frac{\mathrm{E}\left(\mathrm{expit}\left(\beta_0+f_0(X_{0jl})+\delta_{0j}\right)\right)}{\mathrm{E}\left(\mathrm{expit}\left(\beta_0+\beta_1+f_1(X_{1jl})+\delta_{1j}\right)\right)}\neq 1$$

and

$$\frac{\mathrm{E}\left(\mathrm{expit}\left(\beta_0+\beta_1+f_1(X_{1jl})+\delta_{1j}\right)|R_{1jl}=1\right)}{\mathrm{E}\left(\mathrm{expit}\left(\beta_0+f_0(X_{0jl})+\delta_{0j}\right)|R_{0jl}=1\right)}\neq 1$$

even if the two intervention groups have the same missingness mechanism and the same covariate effects. Hence

$$\frac{\mathrm{E}\left(h_1\left(X_{1jl},\delta_{1j}\right)|R_{1jl}=1\right)}{\mathrm{E}\left(h_0\left(X_{0jl},\delta_{0j}\right)|R_{0jl}=1\right)}\neq 1$$

and therefore $\widehat{\mathrm{RR}}^{\mathrm{cr}}_{\mathrm{unadj}}$ is not consistent for true RR. However, under the null hypothesis of no intervention effect ($\beta_1=0$), if the two intervention group have the same missingness mechanism and the same covariate effect, the both ratios of expectations in the right side of equation (D1) equal to one, and hence we have

$$\frac{\mathrm{E}\left(h_1\left(X_{1jl},\delta_{1j}\right)|R_{1jl}=1\right)}{\mathrm{E}\left(h_0\left(X_{0jl},\delta_{0j}\right)|R_{0jl}=1\right)}=1$$

Therefore, if the data generating model has logit link and there is no intervention effect in truth, $\widehat{\mathrm{RR}}^{\mathrm{cr}}_{\mathrm{unadj}}$ is consistent for true RR $=1$ when the two intervention groups have the same missingness and the same covariate effect.

Appendix E

As we defined in equation (8), the adjusted cluster-level estimator of RR using complete records is given by

$$\widehat{\mathrm{RR}}^{\mathrm{cr}}_{\mathrm{adj}}=\frac{\bar{\epsilon}_1^{r(\mathrm{cr})}}{\bar{\epsilon}_0^{r(\mathrm{cr})}}=\frac{\frac{1}{k}\sum_{j=1}^{k}\frac{N^{\mathrm{cr}}_{1j}}{\hat{N}^{\mathrm{cr}}_{1j}}}{\frac{1}{k}\sum_{j=1}^{k}\frac{N^{\mathrm{cr}}_{0j}}{\hat{N}^{\mathrm{cr}}_{0j}}} \tag{E1}$$

where N^{cr}_{ij} and $\hat{N}^{\mathrm{cr}}_{ij}$ are the observed and predicted number of successes for the complete records in the (ij)th cluster.

Assuming the data are generated from the log link model (1) in the main paper, and following the similar argument presented in Appendix B, it can be shown that, in the case of CRA, the numerator of equation (E1) is a consistent estimator of

$$\begin{aligned}\mathrm{E}\left(\frac{N^{\mathrm{cr}}_{1j}}{\hat{N}^{\mathrm{cr}}_{1j}}\right)&=\mathrm{E}\left[\frac{\sum_{l=1}^{m}R_{ijl}\exp\left(\beta_0+\beta_1+f_1(X_{1jl})+\delta_{1j}\right)}{\hat{N}^{\mathrm{cr}}_{1j}(\boldsymbol{X}_{1j},\boldsymbol{R}_{1j})}\right]\\&=\exp\left(\beta_0+\beta_1\right)\mathrm{E}\left[\frac{\exp\left(\delta_{1j}\right)\sum_{l=1}^{m}R_{ijl}\exp\left(f_1(X_{1jl})\right)}{\hat{N}^{\mathrm{cr}}_{1j}(\boldsymbol{X}_{1j},\boldsymbol{R}_{1j})}\right],\end{aligned} \tag{E2}$$

and the denominator of equation (E1) is a consistent estimator of

$$\begin{aligned}\mathrm{E}\left(\frac{N^{\mathrm{cr}}_{0j}}{\hat{N}^{\mathrm{cr}}_{0j}}\right)&=\mathrm{E}\left[\frac{\sum_{l=1}^{m}R_{ijl}\exp\left(\beta_0+f_0(X_{0jl})+\delta_{0j}\right)}{\hat{N}^{\mathrm{cr}}_{0j}(\boldsymbol{X}_{0j},\boldsymbol{R}_{0j})}\right]\\&=\exp\left(\beta_0\right)\mathrm{E}\left[\frac{\exp\left(\delta_{0j}\right)\sum_{l=1}^{m}R_{ijl}\exp\left(f_0(X_{0jl})\right)}{\hat{N}^{\mathrm{cr}}_{0j}(\boldsymbol{X}_{0j},\boldsymbol{R}_{0j})}\right],\end{aligned} \tag{E3}$$

where $\boldsymbol{R}_{ij}$ is the vector of missing outcomes indicators of the (ij)th cluster. The distribution of X (in expectation) is the same between the intervention groups as a consequence of randomisation. The expectations in the right hand side of equations (E2) and (E3) are equal if δ_{0j} and δ_{1j} have common distribution, the missingness mechanism is the same between the intervention groups, and $f_i(X_{ijl}) = f(X_{ijl})$ for $i \in \{0, 1\}$. Hence, we have

$$\widehat{\mathrm{RR}}^{\mathrm{cr}}_{\mathrm{adj}} \to \exp(\beta_1) = \mathrm{RR} \qquad \text{as} \qquad k \to \infty.$$

Therefore, $\widehat{\mathrm{RR}}^{\mathrm{cr}}_{\mathrm{adj}}$ is consistent and, therefore, asymptotically unbiased (as $k \to \infty$) for true RR if (i) the true data generating model is a log link model, (ii) the functional form of the covariates is the same between the intervention groups, (iii) the missingness mechanism is the same between the intervention groups, and (iv) the distribution of random effects is the same between the intervention groups.

The above argument is not true if the data are generated from the logit link model (2) in the main paper with $\beta_1 \neq 0$, and, therefore, $\widehat{\mathrm{RR}}^{\mathrm{cr}}_{\mathrm{adj}}$ is not consistent for true RR ($\neq 1$). However, under the null hypothesis of no intervention effect ($\beta_1 = 0$), the above argument is true if the true data generating model has logit link. Hence $\widehat{\mathrm{RR}}^{\mathrm{cr}}_{\mathrm{adj}}$ is consistent for true RR ($= 1$) as $k \to \infty$.

Appendix F

Table F1 represents the results of the simulation study, explained in the main paper, for RD using cluster-level analyses with full data, CRA and MMI.

Table F1: Average estimates of RD, their average estimated standard errors (SE) and coverage rates for nominal 95% confidence intervals over 1000 simulation runs, using unadjusted cluster-level (CL_U) and adjusted cluster-level (CL_A) analyses with full data, CRA and MMI. Monte Carlo errors for average estimates and average estimated SEs are all less than 0.003 and 0.001, respectively. The true value of RD is 20%.

	k	Average estimate (%)						Average estimated SE						Coverage (%)					
		Full		CRA		MMI		Full		CRA		MMI		Full		CRA		MMI	
		CL_U	CL_A	CL_U	CL_A	CL_U	CL_A	CL_U	CL_A	CL_U	CL_A	CL_U	CL_A	CL_U	CL_A	CL_U	CL_A	CL_U	CL_A
S1	5	20.0	19.9	22.7	22.5	20.2	20.1	0.069	0.051	0.074	0.061	0.074	0.058	93.8	94.3	93.4	90.3	97.3	97.1
	10	20.0	20.1	22.6	22.6	20.1	20.2	0.049	0.037	0.053	0.044	0.053	0.042	95.8	95.1	93.2	91.2	96.5	96.7
	20	20.1	20.1	22.6	22.6	20.2	20.2	0.035	0.027	0.037	0.031	0.037	0.029	95.5	94.0	89.6	86.1	95.5	95.5
	50	20.0	20.0	22.6	22.6	20.1	20.1	0.022	0.017	0.024	0.020	0.023	0.018	95.1	94.8	81.5	75.5	95.2	95.5
S2	5	20.0	20.0	11.7	21.9	19.8	19.8	0.068	0.052	0.083	0.070	0.080	0.066	95.7	94.8	86.8	95.4	98.5	98.8
	10	20.2	20.0	12.0	21.9	20.1	19.9	0.049	0.037	0.059	0.049	0.056	0.045	96.1	95.9	74.4	94.9	97.5	97.3
	20	19.9	19.9	11.7	21.9	20.0	19.9	0.035	0.027	0.042	0.036	0.039	0.032	95.0	94.5	52.2	93.0	94.9	96.2
	50	20.0	20.1	11.8	22.0	20.0	20.1	0.022	0.017	0.027	0.023	0.024	0.020	95.7	94.9	13.6	87.5	95.2	95.7
S3	5	20.2	20.1	19.7	19.6	20.3	20.1	0.068	0.058	0.075	0.067	0.076	0.067	93.8	94.5	93.8	94.1	96.6	97.2
	10	19.9	19.9	19.6	19.6	20.0	20.0	0.050	0.042	0.055	0.048	0.055	0.047	95.7	95.9	95.7	96.1	96.3	96.8
	20	20.0	20.0	19.6	19.6	20.1	20.0	0.036	0.030	0.039	0.034	0.039	0.033	94.6	94.0	94.6	94.1	95.7	95.3
	50	20.0	20.0	19.6	19.6	20.1	20.1	0.023	0.019	0.025	0.022	0.024	0.021	95.4	95.0	95.2	94.7	95.1	94.8
S4	5	20.3	20.2	9.2	17.4	20.0	19.9	0.071	0.058	0.085	0.076	0.086	0.075	94.7	94.0	82.3	94.4	98.6	98.8
	10	20.1	20.1	9.2	17.4	20.2	20.2	0.050	0.042	0.060	0.054	0.059	0.052	93.9	94.5	60.9	92.6	95.9	96.9
	20	19.9	20.0	8.8	17.1	19.9	20.0	0.036	0.030	0.043	0.038	0.041	0.037	95.2	94.1	29.4	89.5	95.5	96.2
	50	20.0	20.0	8.8	17.1	20.0	20.0	0.023	0.019	0.027	0.024	0.026	0.023	95.0	95.7	2.3	80.0	94.8	94.4

Table F2 presents the results of a further simulation study for cluster-level analyses for RD with full data, CRA and MMI. The parameters configuration was the same with the simulation study explained in the main paper except $\beta_1 = 1$ and, in **(S2)** and **(S4)**, $\beta_{2(0)} = 0.5,\ \beta_{2(1)} = 1$.

Table F2: Further simulation results for RD using cluster-level analyses. Average estimates of RD, their average estimated standard errors (SE) and coverage rates for nominal 95% confidence intervals over 1000 simulation runs, using unadjusted cluster-level (CL_U) and adjusted cluster-level (CL_A) analyses with full data, CRA and MMI. The true value of RD is 15%.

	k	Average estimate (%)						Average estimated SE						Coverage (%)					
		Full		CRA		MMI		Full		CRA		MMI		Full		CRA		MMI	
		CL_U	CL_A	CL_U	CL_A	CL_U	CL_A	CL_U	CL_A	CL_U	CL_A	CL_U	CL_A	CL_U	CL_A	CL_U	CL_A	CL_U	CL_A
S1	5	14.9	14.9	16.7	16.6	15.0	15.0	0.071	0.053	0.075	0.063	0.076	0.060	94.5	95.9	94.5	95.2	96.6	98.7
	10	15.1	15.1	16.9	16.8	15.2	15.2	0.050	0.038	0.054	0.045	0.054	0.042	94.2	93.6	93.3	92.3	95.1	94.9
	20	15.1	15.0	16.8	16.7	15.2	15.1	0.036	0.027	0.038	0.032	0.038	0.030	94.5	94.4	92.8	90.8	95.3	94.5
	50	15.1	15.1	16.7	16.7	15.0	15.0	0.023	0.017	0.024	0.020	0.023	0.018	94.6	95.3	88.9	85.2	90.4	90.7
S2	5	14.9	14.9	5.3	15.9	14.9	15.1	0.070	0.052	0.082	0.069	0.083	0.068	94.1	95.2	82.0	95.6	97.6	98.8
	10	14.9	15.1	5.5	16.0	15.1	15.0	0.050	0.038	0.059	0.050	0.058	0.048	94.9	95.1	68.6	95.1	95.5	97.0
	20	15.1	15.1	5.5	16.0	15.0	14.9	0.036	0.027	0.042	0.036	0.041	0.033	94.8	94.5	40.3	94.5	95.4	95.8
	50	15.0	15.0	5.5	16.0	15.0	15.0	0.023	0.017	0.027	0.023	0.025	0.021	94.6	94.3	6.0	94.2	95.8	96.8
S3	5	15.2	15.2	13.2	13.2	15.5	15.4	0.072	0.061	0.078	0.070	0.081	0.071	95.6	96.4	93.7	94.1	97.6	99.1
	10	15.0	15.0	12.9	12.9	15.0	15.1	0.052	0.044	0.056	0.050	0.057	0.050	94.8	94.8	93.5	93.8	96.8	96.8
	20	15.0	15.0	13.0	12.9	15.1	15.1	0.036	0.031	0.039	0.035	0.040	0.035	94.3	93.9	92.7	90.0	95.6	95.4
	50	15.1	15.2	13.0	13.0	15.1	15.2	0.023	0.020	0.025	0.023	0.025	0.022	94.7	96.2	85.4	86.5	94.6	95.4
S4	5	15.1	14.9	1.8	9.5	15.0	14.8	0.072	0.061	0.084	0.076	0.089	0.080	96.0	95.5	73.8	92.8	98.7	99.0
	10	15.1	15.1	1.9	9.8	15.1	15.0	0.051	0.044	0.061	0.055	0.062	0.056	93.6	94.0	45.6	86.3	96.6	97.1
	20	15.1	15.0	1.7	9.7	15.1	15.0	0.036	0.031	0.043	0.039	0.043	0.039	94.4	96.0	15.8	72.2	96.1	97.0
	50	15.0	15.0	1.8	9.8	15.1	15.1	0.023	0.020	0.027	0.025	0.027	0.024	94.6	94.4	0.3	42.6	95.2	95.1

Table F3 shows the results of the simulation study, explained in the main paper, for RR using cluster-level analyses with full data, CRA and MMI.

Table F3: Average estimates of log(RR), their average estimated standard errors (SE) and coverage rates for nominal 95% confidence intervals over 1000 simulation runs, using unadjusted cluster-level (CL_U) and adjusted cluster-level (CL_A) analyses with full data, CRA and MMI. Monte Carlo errors for average estimates and average estimated SEs are all less than 0.005 and 0.001, respectively. The true value of log(RR) is 0.34.

	k	Average estimate						Average estimated SE						Coverage (%)					
		Full		CRA		MMI		Full		CRA		MMI		Full		CRA		MMI	
		CL_U	CL_A	CL_U	CL_A	CL_U	CL_A	CL_U	CL_A	CL_U	CL_A	CL_U	CL_A	CL_U	CL_A	CL_U	CL_A	CL_U	CL_A
S1	5	0.339	0.344	0.461	0.464	0.344	0.348	0.123	0.096	0.159	0.136	0.135	0.110	94.4	93.9	90.5	86.2	97.3	98.0
	10	0.338	0.345	0.456	0.464	0.340	0.348	0.087	0.069	0.114	0.098	0.094	0.078	95.0	94.8	85.7	78.1	96.3	97.4
	20	0.339	0.345	0.456	0.464	0.341	0.348	0.062	0.049	0.080	0.069	0.066	0.054	94.7	93.4	71.6	58.8	95.6	95.3
	50	0.336	0.343	0.453	0.461	0.339	0.346	0.039	0.031	0.051	0.044	0.041	0.034	95.5	95.1	38.2	18.4	95.7	95.5
S2	5	0.339	0.346	0.261	0.515	0.338	0.344	0.122	0.096	0.186	0.161	0.142	0.119	95.8	94.7	94.0	85.8	98.7	98.8
	10	0.341	0.344	0.266	0.514	0.340	0.343	0.087	0.069	0.130	0.112	0.098	0.082	95.7	95.8	92.2	69.1	97.5	97.8
	20	0.336	0.342	0.260	0.512	0.337	0.343	0.062	0.049	0.093	0.081	0.069	0.057	95.4	94.6	88.0	45.4	95.4	96.2
	50	0.337	0.345	0.263	0.516	0.337	0.346	0.039	0.031	0.059	0.052	0.043	0.036	95.7	94.3	76.2	6.1	96.7	95.2
S3	5	0.343	0.342	0.388	0.387	0.347	0.346	0.123	0.107	0.155	0.141	0.140	0.126	93.9	93.6	94.0	93.8	96.7	96.8
	10	0.336	0.338	0.383	0.383	0.338	0.340	0.089	0.077	0.112	0.102	0.099	0.088	95.7	95.8	93.6	94.3	96.5	96.9
	20	0.338	0.339	0.382	0.382	0.339	0.340	0.064	0.055	0.080	0.073	0.070	0.062	94.8	94.3	91.4	90.1	96.0	95.2
	50	0.337	0.339	0.383	0.384	0.339	0.341	0.040	0.035	0.051	0.046	0.044	0.039	95.6	95.2	86.5	84.2	95.8	95.7
S4	5	0.347	0.346	0.200	0.385	0.342	0.341	0.128	0.109	0.186	0.167	0.154	0.138	94.2	94.2	91.8	94.6	98.9	99.0
	10	0.340	0.342	0.198	0.385	0.342	0.344	0.089	0.078	0.130	0.118	0.105	0.095	93.5	94.0	81.8	93.4	95.8	97.2
	20	0.336	0.340	0.189	0.377	0.336	0.339	0.063	0.055	0.092	0.084	0.073	0.066	95.3	94.2	66.2	92.6	95.4	95.6
	50	0.336	0.340	0.189	0.376	0.338	0.340	0.040	0.035	0.058	0.053	0.045	0.041	94.8	96.0	30.4	89.3	94.8	95.4

Table F4 shows the results of a further simulation study for adjusted cluster-level analysis for RR with full data. The parameters configuration was the same with the simulation study explained in the main paper except the variance components parameters for generating the baseline covariate X. We set $\sigma_u^2 = 0.35$, $\sigma_\alpha^2 = 3.20$ and thus we had $\sigma_x^2 = 3.55, \rho_x = 0.9$.

Table F4: Further simulation results for adjusted cluster-level analysis for RR with full data. Average estimates of log(RR), their empirical standard errors (SE), their average estimated SE, and coverage rates for nominal 95% confidence intervals over 1000 simulation runs, using unadjusted cluster-level (CL_U) and adjusted cluster-level (CL_A) analyses with full data. The true value of log(RR) is 0.34.

k	Average estimate		Empirical SE		Average estimated SE		Coverage (%)	
	CL_U	CL_A	CL_U	CL_A	CL_U	CL_A	CL_U	CL_A
5	0.341	0.439	0.336	0.185	0.343	0.208	96.2	94.8
10	0.342	0.460	0.230	0.135	0.234	0.146	95.9	91.4
20	0.341	0.468	0.160	0.097	0.160	0.100	95.7	78.4
50	0.339	0.476	0.100	0.062	0.101	0.063	95.3	38.8
100	0.338	0.477	0.070	0.043	0.070	0.045	95.1	9.4

Table F5 represents the association of the baseline covariates (age, sex, exam score, literacy group and baseline anaemia) of the HALI trial with anaemia at 24 months and with the probability of anaemia outcome at 24 months being missing.

Table F5: Estimates of log odds ratios as measures of association of the baseline covariates with anaemia at 24 months and with the probability of anaemia outcome at 24 months being missing

	Anaemia			Missingness of anaemia		
	Estimate	Std. Error	p-value	Estimate	Std. Error	p-value
Intercept	-1.72	0.81	0.03	-2.10	0.60	0.00
IST (intervention)	0.36	1.10	0.74	-0.27	0.83	0.74
Age (years)	0.07	0.02	< 0.001	0.06	0.02	< 0.001
Sex (male vs female)	-0.04	0.10	0.73	-0.08	0.11	0.48
Exam score	0.00	0.00	0.77	0.00	0.00	0.91
Literacy group	0.06	0.19	0.74	-0.28	0.13	0.03
Baseline anaemia	1.57	0.11	< 0.001	0.09	0.11	0.42
IST: Age *	0.01	0.03	0.62	0.04	0.03	0.12
IST: Sex *	0.10	0.14	0.49	-0.18	0.15	0.24
IST: Exam score *	0.00	0.00	0.59	0.00	0.00	0.62
IST: Literacy group *	0.37	0.26	0.15	0.38	0.19	0.04
IST: Baseline anaemia *	-0.19	0.15	0.19	-0.03	0.15	0.86

* Interaction terms

Part IV

Time-to-Event Outcomes

Chapter 8

Time-to-Event Outcomes

8.1 Introduction

Time-to-event outcomes occur when individuals in the trial are followed until they experience the event of interest or they are censored. In this thesis, we restrict our attention to time to first occurrence of the event of interest. For example, in a trial of reducing fall injury for elderly people, the outcome from each individual is either the time until he/she experiences a fall or the time until he/she is censored.

In Part II and Part III of this thesis, whenever we said that the outcome is missing for an individual we meant that the value of the outcome is entirely unknown for that individual. However, there are some situations where outcome for an individual is neither perfectly known nor entirely unknown. This type of data are known as coarse data. A common source of such data is censoring, which occurs in time-to-event studies when an individual is lost to follow up or outlives the study period. Censoring can be considered as a special case of missing data [13, 44]. The censoring mechanism that gives rise to the censored data can be thought of as the missingness mechanism [44]. We therefore have three kinds of censoring mechanism paralleling MCAR, MAR and MNAR, respectively. Time to event data are said to be censored completely at random (CCAR) when time to censoring is completely independent of time to event. Time to event data are said to be censored at random (CAR) if, conditional on observed data (for example, intervention group or covariates), time to censoring is independent of time to event. Time to event data are said to be censored not at random (CNAR) if time to censoring is dependent of time to event. Censoring mechanisms CCAR and CAR are often referred to as 'non-informative' or 'ignorable' and CNAR often refereed to as 'informative' or 'non-ignorable' [44].

Let T_{ijl} be the time to event and C_{ijl} be the time to censoring for the lth ($l = 1, 2, \ldots, m_{ij}$) individual in the jth ($j = 1, 2, \ldots, k_i$) cluster of the ith ($i = 0, 1$) intervention group, where $i = 0$ corresponds to control group and $i = 1$ corresponds to active intervention group. Then, for each individual, we observe the follow-up time $Y_{ijl} = \min(T_{ijl}, C_{ijl})$ and a event indicator Δ_{ijl}, where $\Delta_{ijl} = 1$ if $T_{ijl} < C_{ijl}$ and $\Delta_{ijl} = 0$ otherwise. Also let X_{ijl} be an individual-level baseline covariate for the lth individual in the (ij)th cluster. For simplicity, we assume here that we have only one baseline

covariate, though in practice, X can be a vector of covariates, some of which are at the individual-level and some of which are at the cluster-level. For convenience, we assume that both control and intervention groups have the same number of clusters ($k_i = k$) and constant cluster size ($m_{ij} = m$) across the intervention groups.

For time-to-event data, the rate ratio (RaR) or hazard ratio is usually used as the measure of intervention effect [5]. In the literature, the two broad approaches for estimating RaR in CRTs are cluster-level analysis and individual-level analysis. As far as we are aware, there has been little work done to investigate the consistency of the cluster-level analysis methods under different scenarios, for example, with or without censored data.

An alternative way to measure the intervention effect could be to compare the survival functions between the intervention groups. One advantages of using survival functions to quantify the intervention effect is that this approach doesn't rely on any assumptions that the rate/hazard is constant over time. Survival functions can be estimated using the Kaplan-Meier (KM) estimator, assuming that censoring times and survival times are independent, but the standard error of these estimates need to be adjusted for the clustered structure of the data. Greenwood's formula is often used to estimate the variance of KM estimates assuming the observations are statistically independent. In our CRT setting, these variances need to be adjusted to acknowledge the clustered structure of the data. Williams (1995) [38] derived a variance estimator for KM estimates considering the observations are not statistically independent. Our impression is that this methodology has not been widely used, particularly in the setting of CRTs.

We now describe the data generating mechanism assumed in the remainder of the chapter. We are going to assume that the rate or hazard is constant over time, which is in contrast to what is done in non-clustered randomised trials, where the Cox model usually used, which doesn't assume constant hazards. Suppose $T_{ijl} \sim \text{Exp}(\lambda_{ijl})$, where the rate

$$\lambda_{ijl} = \delta_{ij} \exp(\beta_0 + \beta_1 i + f_i(X_{ijl})) \tag{8.1}$$

with β_0 a constant, β_1 is the intervention effect, $f_i(X_{ijl})$ is a function of the baseline covariate X_{ijl} in the ith intervention group, and δ_{ij} is the random effect for the (ij)th cluster. In order to separate the baseline rate from the overall random effect, the mean of the random effect is typically constrained to unity.

The true RaR can then be defined as

$$\begin{aligned} \text{RaR} &= \frac{\text{E}_{jl}(\lambda_{1jl})}{\text{E}_{jl}(\lambda_{0jl})} \\ &= \frac{\text{E}_{jl}[\delta_{1j} \exp(\beta_0 + \beta_1 + f_1(X_{1jl}))]}{\text{E}_{jl}[\delta_{0j} \exp(\beta_0 + f_0(X_{0jl}))]} \\ &= \exp(\beta_1) \frac{\text{E}_{jl}[\delta_{1j} \exp(f_1(X_{1jl}))]}{\text{E}_{jl}[\delta_{0j} \exp(f_0(X_{0jl}))]} \end{aligned} \tag{8.2}$$

Since the distribution of X_{ijl} is the same across the intervention groups by randomisation, the expectations in the numerator and denominator of equation (8.2) will be the same if $f_i(X_{ijl}) = f(X_{ijl}), i \in (0, 1)$. Under these assumptions, we have $\text{RaR} = \exp(\beta_1)$.

This chapter is organised as follows. Section 8.2 describes the cluster-level analysis methods, investigates under which conditions these methods are consistent for estimating RaR, and presents a simulation study. Section 8.3 describes the shared frailty model, an individual-level analysis method, and investigates its performance through simula-

tion. Section 8.4 explains the Kaplan-Meier estimate of survival function and Williams' approach for estimating the standard errors for KM estimates considering the clustered structure of the data and presents a simulation study. We summarise the chapter in Section 8.5.

8.2 Cluster-level analysis

We now explain briefly, how to define and conduct unadjusted and adjusted cluster level analyses for time-to-event outcomes.

8.2.1 Unadjusted cluster-level analysis

Similar to what was done with continuous and binary outcomes, a relevant summary measure of outcomes is calculated for each cluster in the first stage of analysis. For time-to event data, the cluster-level rate of the event of interest is usually used as the summary measure for each cluster [5]. Let r_{ij} be the observed rate of the event of interest in the jth cluster of the ith intervention group. Then RaR is estimated as

$$\widehat{\mathrm{RaR}}_{\mathrm{unadj}} = \frac{\bar{r}_1}{\bar{r}_0}$$

where $\bar{r}_i$ is the mean of the cluster-specific event rates in the ith intervention group. Then in the second stage, a test of the hypothesis $\log(\mathrm{RaR}_{\mathrm{unadj}}) = 0$ is performed using a standard two independent sample $t-$test with DF $2k-2$, where $\widehat{\mathrm{Var}}\big(\log(\widehat{\mathrm{RaR}}_{\mathrm{unadj}})\big)$

can be estimated by

$$\widehat{\text{Var}}\big(\log(\widehat{\text{RaR}}_{\text{unadj}})\big) = \frac{s_0^2}{k\bar{r}_0^2} + \frac{s_1^2}{k\bar{r}_1^2} \quad \text{with} \quad s_i^2 = \frac{\sum_{j=1}^{k}(r_{ij} - \bar{r}_i)^2}{k-1}$$

A 95% confidence interval (CI) for the $\log(\text{RaR}_{\text{unadj}})$ can be obtain based on $t-$distribution as

$$\log\left(\widehat{\text{RaR}}_{\text{unadj}}\right) \pm t_{2k-2,0.025} \times \sqrt{\widehat{\text{Var}}\big(\log(\widehat{\text{RaR}}_{\text{unadj}})\big)}$$

One can then easily obtain a 95% CI for the $\text{RaR}_{\text{unadj}}$ by dividing and multiplying the $\widehat{\text{RaR}}_{\text{unadj}}$ by

$$\exp\left(t_{2k-2,0.025} \times \sqrt{\widehat{\text{Var}}\big(\log(\widehat{\text{RaR}}_{\text{unadj}})\big)}\right).$$

We now investigate the consistency of $\widehat{\text{RaR}}_{\text{unadj}}$. First we consider the case of no censoring in the data and then we consider the case of censoring in the data.

No censoring: Recall that the number of individuals in each cluster is m. Since there is no censoring, each individual will have an event observed. Hence, the observed rate of event for the (ij)th cluster is given by

$$r_{ij} = \frac{m}{\sum_{l=1}^{m} T_{ijl}}.$$

The RaR is then estimated as

$$\widehat{\text{RaR}}_{\text{unadj}} = \frac{\bar{r}_1}{\bar{r}_0} = \frac{(1/k)\sum_{j=1}^{k} r_{1j}}{(1/k)\sum_{j=1}^{k} r_{0j}} \tag{8.3}$$

Now for ith intervention group

$$\begin{aligned}\bar{r}_i = k^{-1}\sum_{j=1}^{k} r_{ij} &\xrightarrow[k\to\infty]{prob} \mathrm{E}(r_{ij}) \\ &= \mathrm{E}\left(\frac{m}{\sum_{l=1}^{m} T_{ijl}}\right) \\ &= \mathrm{E}\left(\frac{1}{(1/m)\sum_{l=1}^{m} T_{ijl}}\right) \\ &\approx \frac{1}{(1/m)\sum_{l=1}^{m}\mathrm{E}(T_{ijl})}, \end{aligned} \tag{8.4}$$

using delta method with m is large. Now

$$\begin{aligned}\mathrm{E}(T_{ijl}) &= \mathrm{E}_{jl}[\mathrm{E}_{jl}(T_{ijl}|\lambda_{ijl})] \\ &= \mathrm{E}_{jl}(\lambda_{ijl}^{-1}) \text{ since } T_{ijl} \sim \mathrm{Exp}(\lambda_{ijl}) \\ &= \mathrm{E}_{jl}[\delta_{ij}^{-1}\exp(-\beta_0 - \beta_1 i - f_i(X_{ijl}))] \\ &= \exp(-\beta_0 - \beta_1 i)\mathrm{E}_{jl}[\delta_{ij}^{-1}\exp(-f_i(X_{ijl}))]\end{aligned}$$

Plugging this result into equation (8.4), we have

$$\bar{r}_i = k^{-1}\sum_{j=1}^{k} r_{ij} \xrightarrow[(k,m)\to\infty]{prob} \frac{\exp(\beta_0 + \beta_1 i)}{\mathrm{E}_{jl}[\delta_{ij}^{-1}\exp(-f_i(X_{ijl}))]}$$

Hence

$$\widehat{\mathrm{RaR}}_{\text{unadj}} = \frac{\bar{r}_1}{\bar{r}_0} \xrightarrow[(k,m)\to\infty]{prob} \exp(\beta_1)\frac{\mathrm{E}_{jl}[\delta_{0j}^{-1}\exp(-f_0(X_{0jl}))]}{\mathrm{E}_{jl}[\delta_{1j}^{-1}\exp(-f_1(X_{1jl}))]}. \tag{8.5}$$

Since the distribution of X is the same across the intervention groups by randomisation, the expectations in the numerator and denominator in equation (8.5) will cancel out if $f_i(X_{ijl}) = f(X_{ijl}), i \in (0, 1)$. Under these assumption, we have

$$\widehat{\text{RaR}}_{\text{unadj}} \xrightarrow[(k,m)\to\infty]{prob} \exp(\beta_1) = \text{RaR}$$

Therefore, with no censoring, $\widehat{\text{RaR}}$ is consistent as $(m, k) \to \infty$ if the functional form of the covariates is the same between the intervention groups in the data generating model for the outcome. It can also be concluded here that in the case of no covariates effects, in which case the event rate is the same for every individual within the intervention groups, $\widehat{\text{RaR}}$ is also consistent.

Censored data: In the case of censored data, RaR is estimated as

$$\widehat{\text{RaR}}_{\text{unadj}} = \frac{(1/k)\sum_{j=1}^{k} r_{1j}}{(1/k)\sum_{j=1}^{k} r_{0j}} \tag{8.6}$$

where

$$r_{ij} = \frac{\sum_{l=1}^{m} \Delta_{ijl}}{\sum_{l=1}^{m} Y_{ijl}}.$$

Consider cluster j in intervention group i. Then the survival function in the (ij)th cluster is defined as

$$\begin{aligned} S_{ij}(t) &= \mathrm{E}_l[\mathbb{1}(T_{ijl} > t)] \\ &= \mathrm{E}_l[\mathrm{E}_l[\mathbb{1}(T_{ijl} > t)|\lambda_{ijl}]] \\ &= \mathrm{E}_l[S_{ij}(t|\lambda_{ijl})] \\ &= \mathrm{E}_l[\exp(-\lambda_{ijl}t)] \end{aligned} \tag{8.7}$$

- **Condition 1**: If $\mathrm{Var}_l(\lambda_{ijl})$ is small (and so X_{ijl} has small effect), we can approximate the survival function from equation (8.7) using delta method. We have

$$S_{ij}(t) \approx \exp[-t\mathrm{E}_l(\lambda_{ijl})] \tag{8.8}$$

  which is the survival function of an exponential distribution with rate $\mathrm{E}_l(\lambda_{ijl})$.

- **Condition 2**: If $\lambda_{ijl}t$ is small (which is possible when either study period is short or rates are small), using the Taylor expansion of the exponential function, we can write from equation (8.7)

$$\begin{aligned} S_{ij}(t) &\approx \mathrm{E}_l[1 - \lambda_{ijl}t] \\ &= 1 - t\mathrm{E}_l(\lambda_{ijl}) \\ &\approx \exp[-t\mathrm{E}_l(\lambda_{ijl})], \end{aligned} \tag{8.9}$$

  which will be valid when $t\mathrm{E}_l(\lambda_{ijl})$ is small. Under these assumptions, $S_{ij}(t)$ is the survival function of an exponential distribution again with rate $\mathrm{E}_l(\lambda_{ijl})$.

Hence, under either of these conditions assuming censoring is independent within (ij)th cluster, we have

$$r_{ij} \xrightarrow[m\to\infty]{prob} \mathrm{E}_l(\lambda_{ijl}) = \delta_{ij} \exp(\beta_0 + \beta_1 i)\mathrm{E}_l[\exp(f_i(X_{ijl}))] \tag{8.10}$$

since r_{ij} is the MLE for common rate for a sample of independent individuals with exponentially distributed event times. Then

$$\begin{aligned}\widehat{\mathrm{RaR}}_{\mathrm{unadj}} &\xrightarrow[(k,m)\to\infty]{prob} \frac{\mathrm{E}_j(r_{1j})}{\mathrm{E}_j(r_{0j})} \\ &= \exp(\beta_1)\frac{\mathrm{E}_j[\delta_{1j}\mathrm{E}_l[\exp(f_1(X_{1jl}))]]}{\mathrm{E}_j[\delta_{0j}\mathrm{E}_l[\exp(f_0(X_{0jl}))]]}.\end{aligned} \tag{8.11}$$

Again, since the distribution of X is the same across the intervention groups by randomisation, the expectations in the numerator and denominator in equation (8.11) will the same if $f_i(X_{ijl}) = f(X_{ijl}), i \in (0, 1)$ (i.e. if the rates depend on X in both groups according to the same functional form). Under these assumptions,

$$\widehat{\mathrm{RaR}}_{\mathrm{unadj}} \xrightarrow[(k,m)\to\infty]{prob} \exp(\beta_1) = \mathrm{RaR}$$

Therefore, in the case with censored data, $\widehat{\mathrm{RaR}}$ is consistent as $(k, m) \to \infty$ if (i) either the covariate has small effect or event rates are small, (ii) the functional form of the covariates is the same between the intervention groups in the data generating model for the outcome, and (iii) censoring is independent within each cluster.

8.2.2 Adjusted cluster-level analysis

In adjusted cluster-level analysis, an individual-level regression analysis of the outcome of interest is carried out at the first stage of analysis ignoring the clustering of the data, which incorporates all covariates into the regression model except intervention indicator[5, 7]. A standard Poisson regression model is usually fitted for time-to-event data, which assumes

$$\log(\lambda_{ijl}) = \alpha_1 + \alpha_2 X_{ijl} \tag{8.12}$$

Let N_{ij} and $\hat{N}_{ij}$ be the observed and predicted number of event in the (ij)th cluster, respectively. After fitting the model (8.12), $\hat{N}_{ij}$ is calculated as

$$\hat{N}_{ij} = \sum_{l=1}^{m} Y_{ijl}\hat{\lambda}_{ijl} = \sum_{l=1}^{m} Y_{ijl} \times \exp(\hat{\alpha}_1 + \hat{\alpha}_2 X_{ijl}) \tag{8.13}$$

Then the ratio-residual for each cluster is calculated as

$$\epsilon_{ij} = \frac{N_{ij}}{\hat{N}_{ij}}$$

The adjusted RaR is then estimated as

$$\widehat{\mathrm{RaR}}_{\mathrm{adj}} = \frac{\bar{\epsilon}_1}{\bar{\epsilon}_0} = \frac{\frac{1}{k}\sum_{j=1}^{k} \frac{N_{1j}}{\hat{N}_{1j}}}{\frac{1}{k}\sum_{j=1}^{k} \frac{N_{0j}}{\hat{N}_{0j}}} \tag{8.14}$$

In the second stage, a test of the hypothesis $\log(\mathrm{RaR}_{\mathrm{adj}}) = 0$ is performed using a standard two independent sample $t-$test with degrees of freedom (DF) $2k-2$, where $\widehat{\mathrm{Var}}\big(\log(\widehat{\mathrm{RaR}}_{\mathrm{adj}})\big)$ can be estimated by

$$\widehat{\mathrm{Var}}\big(\log(\widehat{\mathrm{RaR}}_{\mathrm{adj}})\big) = \frac{s^2_{\epsilon 0}}{k\bar{\epsilon}^2_0} + \frac{s^2_{\epsilon 1}}{k\bar{\epsilon}^2_1} \quad \text{with} \quad s^2_{\epsilon i} = \frac{\sum_{j=1}^{k}(\epsilon_{ij} - \bar{\epsilon}_i)^2}{k-1}$$

A 95% confidence interval (CI) for the $\log(\mathrm{RaR}_{\mathrm{adj}})$ can be calculated as

$$\log\left(\widehat{\mathrm{RaR}}_{\mathrm{adj}}\right) \pm t_{2k-2,0.025} \times \sqrt{\widehat{\mathrm{Var}}\big(\log(\widehat{\mathrm{RaR}}_{\mathrm{adj}})\big)}$$

One can then easily obtain a 95% CI for the $\mathrm{RaR}_{\mathrm{adj}}$ by dividing and multiplying the $\widehat{\mathrm{RaR}}_{\mathrm{adj}}$ by

$$\exp\left(t_{2k-2,0.025} \times \sqrt{\widehat{\mathrm{Var}}\big(\log(\widehat{\mathrm{RaR}}_{\mathrm{adj}})\big)}\right).$$

We now investigate the consistency of $\widehat{\mathrm{RaR}}_{\mathrm{adj}}$. First we consider the case of no censoring in the data and then we consider censoring in the data.

No censoring: Recall that each cluster has m individuals. Since in the case of no censoring each individual will have an event, the observed number of event in each cluster is $N_{ij} = m$, and the predicted number of event is $\hat{N}_{ij} = \sum_{l=1}^{m} T_{ijl}\hat{\lambda}_{ijl}$. Again, since $\hat{\lambda}_{ijl}$ is a function of X_{ijl} only, replacing $\hat{\lambda}_{ijl}$ by $h(X_{ijl})$ for convenience, we have

$$\hat{N}_{ij} = \sum_{l=1}^{m} T_{ijl}h(X_{ijl}).$$

The numerator of (8.14) is a consistent estimator of

$$\begin{aligned}\mathrm{E}\left(\frac{N_{1j}}{\hat{N}_{1j}}\right) &= \mathrm{E}\left(\frac{m}{\sum_{l=1}^{m} T_{1jl}\hat{\lambda}_{1jl}}\right) \\ &= \mathrm{E}\left(\frac{1}{(1/m)\sum_{l=1}^{m} T_{1jl}\ h(X_{1jl})}\right),\end{aligned}$$

Then using the delta method assuming large m, we have

$$\begin{aligned}\mathrm{E}\left(\frac{N_{1j}}{\hat{N}_{1j}}\right) &\approx \frac{1}{\mathrm{E}[(1/m)\sum_{l=1}^{m} T_{1jl}\ h(X_{1jl})]} \\ &= \left(\mathrm{E}\left[\frac{1}{m}\sum_{l=1}^{m} T_{1jl}\ h(X_{1jl})\right]\right)^{-1} \\ &= (\mathrm{E}[T_{1jl}h(X_{1jl})])^{-1} \qquad (8.15)\end{aligned}$$

Now

$$\begin{aligned}\mathrm{E}[T_{1jl}h(X_{1jl})] &= \mathrm{E}\left[\mathrm{E}\left(T_{1jl}\ h(X_{1jl})|X_{1jl},\delta_{1j}\right)\right] \\ &= \mathrm{E}\left[h(X_{1jl})\mathrm{E}\left(T_{1jl}|X_{1jl},\delta_{1j}\right)\right] \\ &= \mathrm{E}[h(X_{1jl})\lambda_{1jl}^{-1}] \\ &= \mathrm{E}[h(X_{1jl})\delta_{1j}^{-1}\exp(-\beta_0-\beta_1-f_1(X_{1jl}))] \\ &= \exp(-\beta_0-\beta_1)\mathrm{E}[\delta_{1j}^{-1}\ h(X_{1jl})\exp(-f_1(X_{1jl}))]\end{aligned}$$

Plugging this result into equation (8.15), we have

$$\mathrm{E}\left(\frac{N_{1j}}{\hat{N}_{1j}}\right) = \exp(\beta_0+\beta_1)\big(\mathrm{E}[\delta_{1j}^{-1}\ h(X_{1jl})\exp(-f_1(X_{1jl}))]\big)^{-1} \qquad (8.16)$$

Analogously, it can be shown that the denominator of equation (8.14) is a consistent estimator of

$$\mathrm{E}\left(\frac{N_{0j}}{\hat{N}_{0j}}\right) = \exp(\beta_0)\big(\mathrm{E}[\delta_{0j}^{-1} h(X_{0jl}) \exp(-f_0(X_{0jl}))]\big)^{-1} \tag{8.17}$$

Since the distribution of X is in expectation the same between the intervention groups as a consequence of randomisation, the expectations in the right hand sides of equations (8.16) and (8.17) are equal. Under all these assumptions, we have

$$\widehat{\mathrm{RaR}}_{\mathrm{adj}} \xrightarrow[(k,m)\to\infty]{prob} \exp(\beta_1) = \mathrm{RaR}.$$

Note that this proof has not needed to assume that the first stage regression model is correctly specified.

Censored data: The numerator of (8.14) is a consistent estimator of

$$\begin{aligned}
\mathrm{E}\left(\frac{N_{1j}}{\hat{N}_{1j}}\right) &= \mathrm{E}\left[\mathrm{E}\left(\frac{N_{1j}}{\hat{N}_{1j}}\middle|\boldsymbol{Y}_{1j}, \boldsymbol{X}_{1j}, \delta_{1j}\right)\right] \\
&= \mathrm{E}\left[\frac{\mathrm{E}\left(N_{1j}|\boldsymbol{Y}_{1j}, \boldsymbol{X}_{1j}, \delta_{1j}\right)}{\sum_{l=1}^{m} Y_{1jl}\hat{\lambda}_{1jl}}\right], \text{ assuming independent censoring} \\
&= \mathrm{E}\left[\frac{\sum_{l=1}^{m} Y_{1jl}\lambda_{1jl}}{\sum_{l=1}^{m} Y_{1jl}\hat{\lambda}_{1jl}}\right] \\
&= \mathrm{E}\left[\frac{\sum_{l=1}^{m} Y_{1jl}\,\delta_{1j}\exp\left(\beta_0+\beta_1+f_1(X_{1jl})\right)}{\sum_{l=1}^{m} Y_{1jl}\exp\left(\hat{\alpha}_1+\hat{\alpha}_2 X_{1jl}\right)}\right] \\
&= \frac{\exp\left(\beta_0+\beta_1\right)}{\exp\left(\hat{\alpha}_1\right)}\ \mathrm{E}\left[\frac{\delta_{1j}\sum_{l=1}^{m} Y_{1jl}\exp\left(f_1(X_{1jl})\right)}{\sum_{l=1}^{m} Y_{1jl}\exp\left(\hat{\alpha}_2 X_{1jl}\right)}\right]
\end{aligned}$$

where $\boldsymbol{Y}_{1j}$ and $\boldsymbol{X}_{1j}$ are the vectors of Y_{1jl} and X_{1jl} values, respectively, for the j cluster of the intervention group. Suppose $f_i(X_{ijl}) = f(X_{ijl}) = \beta_2 X_{ijl},\ i \in (0,1)$. Then

$$\mathrm{E}\left(\frac{N_{1j}}{\hat{N}_{1j}}\right) = \frac{\exp\left(\beta_0 + \beta_1\right)}{\exp\left(\hat{\alpha}_1\right)}\ \mathrm{E}\left[\frac{\delta_{1j}\sum_{l=1}^{m} Y_{1jl}\exp\left(\beta_2 X_{1jl}\right)}{\sum_{l=1}^{m} Y_{1jl}\exp\left(\hat{\alpha}_2 X_{1jl}\right)}\right] \tag{8.18}$$

The first stage model (8.12) incorporates only the covariate X_{ijl} and not the intervention indicator. This model possesses the collapsibility property because it assumes log-link and X_{ijl} is independent of the intervention indicator. Hence, $\hat{\alpha}_2$ is consistent for β_2 assuming censoring is independent conditional on X, and therefore

$$\mathrm{E}\left[\frac{\delta_{1j}\sum_{l=1}^{m} Y_{1jl}\exp\left(\beta_2 X_{1jl}\right)}{\sum_{l=1}^{m} Y_{1jl}\exp\left(\hat{\alpha}_2 X_{1jl}\right)}\right] = \mathrm{E}(\delta_{1j}) \tag{8.19}$$

This result is true for any form of $f(X_{ijl})$, provided that the analyst correctly models the dependence on X in the first stage model.

Plugging equation (8.19) into equation (8.18), we have

$$\mathrm{E}\left(\frac{N_{1j}}{\hat{N}_{1j}}\right) = \frac{\exp\left(\beta_0 + \beta_1\right)}{\exp\left(\hat{\alpha}_1\right)}\ \mathrm{E}(\delta_{1j}).$$

Using a similar argument, it can be shown that the denominator of (8.14) is a consistent estimator of

$$\mathrm{E}\left(\frac{N_{0j}}{\hat{N}_{0j}}\right) = \frac{\exp(\beta_0)}{\exp(\hat{\alpha}_1)}\ \mathrm{E}(\delta_{0j}). \tag{8.20}$$

Now, from equation (8.14), we can write

$$\begin{aligned}\widehat{\mathrm{RaR}}_{\mathrm{adj}} &\xrightarrow[k\to\infty]{prob} \frac{\mathrm{E}\left(\frac{N_{1j}}{\hat{N}_{1j}}\right)}{\mathrm{E}\left(\frac{N_{0j}}{\hat{N}_{0j}}\right)} \\ &= \frac{\exp\left(\beta_0+\beta_1\right)}{\exp\left(\hat{\alpha}_1\right)} \times \frac{\exp\left(\hat{\alpha}_1\right)}{\exp\left(\beta_0\right)} \times \frac{\mathrm{E}\left(\delta_{1j}\right)}{\mathrm{E}\left(\delta_{0j}\right)} \\ &= \exp(\beta_1) \times \frac{\mathrm{E}\left(\delta_{1j}\right)}{\mathrm{E}\left(\delta_{0j}\right)} \\ &= \exp(\beta_1) = \mathrm{RaR},\end{aligned}$$

since δ_{1j} and δ_{01} have the same mean by assumption. Therefore, in the case of censored data, adjusted cluster-level estimator of RaR is consistent if the rates depend on covariates in both groups according to the same functional form ($f_i(X_{ijl}) = f(X_{ijl}), i \in (0,1)$) and the first stage regression model correctly specifies such dependence.

8.2.3 Simulation study I

A simulation study was conducted to investigate the consistency of the cluster-level analysis methods for estimating RaR. We considered three different scenarios: no censoring, only administrative censoring, and only random censoring that depends on baseline covariate values.

Data generation and analysis: Data were generated using the model in equation (8.1) with $f_i(X_{ijl}) = \beta_2 X_{ijl}$, where β_2 is the effect of covariate X_{ijl}. For each individual in the study, a value of the covariate X_{ijl} was generated using the model

$$X_{ijl} = \alpha_{ij} + u_{ijl}$$

where α_{ij} is the (ij)th cluster effect on X and u_{ijl} is the individual-level error on X. We assumed that $\alpha_{ij} \sim \mathcal{N}(\mu_x, \sigma_\alpha^2)$ independently of $u_{ijl} \sim \mathcal{N}(0, \sigma_u^2)$, where μ_x is the mean of X, σ_α^2 and σ_u^2 are the between-cluster and within-cluster variance of X, respectively. The total variance of X can be written as $\sigma_x^2 = \sigma_\alpha^2 + \sigma_u^2$ and thus the ICC of X is $\rho_x = \sigma_\alpha^2/\sigma_x^2$. Then we generated event times for each individual as $T_{ijl} \sim \text{Exp}(\lambda_{ijl})$, where

$$\lambda_{ijl} = \delta_{ij} \exp(\beta_0 + \beta_1 i + \beta_2 X_{ijl}) \tag{8.21}$$

with $\delta_{ij} \sim \text{Gamma}\,(\text{shape} = 1/\theta_1, \text{rate} = 1/\theta_1)$. The non-parametric intraclass correlation coefficient for T_{ijl} was calculated as $\rho_T = \theta_1/(2+\theta_1)$. For generating individual-level covariate values X_{ijl}, we chose $\mu_x = 6, \sigma_x^2 = 1$ and $\rho_x = 0.05$; and thus we had $\sigma_\alpha^2 = 0.05$ and $\sigma_u^2 = 0.95$. We set the parameter $\beta_1 = -0.35$, which corresponds to true RaR=0.70, and $\beta_2 = 1$. The value of the parameter for generating random effects was fixed as $\theta_1 = (2/9, 2/19)$ so that we had $\rho_T = (0.1, 0.05)$, respectively. We fixed $\beta_0 = -8.0$ for no censoring. For only administrative censoring, we varied $\beta_0 = (-10.5, -8.0)$ and the length of the study $\tau = (3, 5, 7, 10, 15)$ to have the proportions of the event of interest as small to moderate or moderate to high in the intervention groups.

For random censoring, we generated random censoring times for each individual as $C_{ijl} \sim \text{Exp}(\phi_{ijl})$, where

$$\phi_{ijl} = \omega_{ij} \exp(\psi_0 + \psi_1 X_{ijl})$$

with $\omega_{ij} \sim \text{Gamma}\,(\text{shape} = 1/\theta_2, \text{rate} = 1/\theta_2)$. We set $\psi_0 = -1.5$, $\psi_1 = 0.5$ and $\theta_2 = 2/19$ and thus we had $\rho_C = 0.05$. We varied $\beta_0 = (-8.5, -6.5, -5.5, -4.5, -3.5, -3.0)$ to have low to high proportions of event. In this simulation study, we also varied the number of clusters in each intervention group as $k = (5, 10, 20, 30)$ and fixed the cluster size $m = 100$. We estimated log(RaR) from each generated data set as a measure of intervention effect.

Recall that, in the case of adjusted cluster-level analysis with no censoring, we showed analytically (Section 8.2.2) that the first stage regression model need not to be correctly specified to get consistent estimate of RaR. To show this empirically in an additional set of analyses, we generated event times for each individual as $T_{ijl} \sim \text{Exp}(\lambda_{ijl})$, where

$$\lambda_{ijl} = \delta_{ij} \exp(\beta_0 + \beta_1 i + \beta_2 X_{ijl} + \beta_3 \sqrt{X_{ijl}})$$

but we fit the first stage model as

$$\log(\lambda_{ijl}) = \alpha_1 + \alpha_2 X_{ijl}$$

Results: The average estimates of log(RaR), their average estimated standard errors (aveSE), empirical standard errors (empSE), and coverage rates for nominal 95% confidence interval are presented over 1000 simulation runs with $\rho_T = 0.1$ for each of the three scenarios: no censoring, only administrative censoring and only random censoring.

For unadjusted cluster-level analysis, Table 8.1 shows the results for no censoring, Table 8.2 and Table 8.3 show the results for administrative censoring considering low to moderate and moderate to high proportions of event, respectively, and Table 8.4 shows the results for only random censoring. In the case of no censoring, the average estimates of log(RaR) were very close to the true value with coverage rates close to the nominal rate (see Table 8.1). In the cases of censored data (either administrative or random censoring), the average estimates of log(RaR) were close to the true value of RaR with good coverage rates when the proportions of event were small (see Table 8.2 with $\tau = 3, 5$ and Table 8.4). In contrast, as the proportions of event in the intervention groups went high (see Table 8.3 and Table 8.4), the average estimates of log(RaR) went away from the true RaR. These empirical results support our derived analytical results in Section 8.2.1 for estimating RaR using unadjusted cluster-level analysis. We observed qualitatively similar results under $\rho_T = 0.05$ and are not presented.

Table 8.1: Simulation results for the unadjusted cluster-level analysis considering no censoring. Average estimates of log(RaR), their average estimated standard errors (aveSE), empirical standard errors (empSE), and coverage (Cov) rates for nominal 95% confidence interval over 1000 simulation runs are presented. The true log(RaR) is -0.35.

k	Estimate	aveSE	empSE	Cov (%)
5	-0.349	0.332	0.356	95.2
10	-0.355	0.246	0.250	94.5
20	-0.352	0.177	0.184	94.6
30	-0.354	0.146	0.147	95.0

Table 8.2: Simulation results for the unadjusted cluster-level analysis considering only administrative censoring with low to moderate proportions of event. Average estimates of log(RaR), their average estimated standard errors (aveSE), empirical standard errors (empSE), and coverage (Cov) rates for nominal 95% confidence interval over 1000 simulation runs are presented. The true log(RaR) is -0.35.

k	τ	Proportions of event		Estimate	aveSE	empSE	Cov (%)
		Control	Intervention				
5	3	0.051	0.037	-0.358	0.420	0.458	94.8
	5	0.080	0.059	-0.347	0.370	0.406	93.9
	7	0.107	0.079	-0.343	0.348	0.362	94.5
	10	0.145	0.109	-0.338	0.329	0.347	94.4
	15	0.199	0.155	-0.326	0.315	0.333	94.4
10	3	0.051	0.037	-0.334	0.311	0.314	95.5
	5	0.081	0.059	-0.327	0.273	0.284	94.9
	7	0.108	0.080	-0.323	0.258	0.272	93.1
	10	0.144	0.109	-0.307	0.246	0.250	94.0
	15	0.202	0.153	-0.317	0.231	0.234	94.8
20	3	0.051	0.036	-0.336	0.221	0.222	95.1
	5	0.080	0.058	-0.328	0.198	0.198	95.7
	7	0.108	0.080	-0.326	0.186	0.191	94.3
	10	0.145	0.109	-0.316	0.176	0.174	95.2
	15	0.199	0.153	-0.304	0.167	0.165	94.9
30	3	0.051	0.036	-0.343	0.181	0.182	95.6
	5	0.081	0.059	-0.335	0.163	0.165	94.1
	7	0.108	0.080	-0.324	0.153	0.155	95.2
	10	0.146	0.108	-0.321	0.144	0.147	94.4
	15	0.201	0.152	-0.315	0.135	0.131	93.8

For adjusted cluster-level analysis when the first stage model is correctly specified, Table 8.5 shows the results for no censoring, Table 8.6 and Table 8.7 show the results for administrative censoring considering low to moderate and moderate to high proportions of event, respectively, and Table 8.8 shows the results for only random censoring. The average estimates of log(RaR) were very close to the true value of RaR with good coverage rates regardless of whether there was censoring or not. These empirical results support our derived analytical results in Section 8.2.2 for estimating RaR using adjusted

Table 8.3: Simulation results for the unadjusted cluster-level analysis considering only administrative censoring with moderate to high proportions of event. Average estimates of log(RaR), their average estimated standard errors (aveSE), empirical standard errors (empSE), and coverage (Cov) rates for nominal 95% confidence interval over 1000 simulation runs are presented. The true log(RaR) is -0.35.

k	τ	Proportions of event		Estimate	aveSE	empSE	Cov (%)
		Control	Intervention				
5	3	0.368	0.294	-0.294	0.293	0.314	94.8
	5	0.492	0.405	-0.295	0.276	0.298	93.2
	7	0.577	0.490	-0.279	0.272	0.273	95.4
	10	0.662	0.577	-0.275	0.267	0.287	94.3
	15	0.750	0.673	-0.281	0.265	0.298	93.1
10	3	0.369	0.295	-0.292	0.209	0.214	93.2
	5	0.489	0.407	-0.280	0.201	0.200	93.6
	7	0.574	0.488	-0.278	0.196	0.199	94.5
	10	0.661	0.575	-0.289	0.193	0.200	93.1
	15	0.751	0.674	-0.272	0.195	0.207	93.2
20	3	0.368	0.296	-0.289	0.150	0.154	92.2
	5	0.491	0.406	-0.290	0.143	0.144	93.5
	7	0.573	0.488	-0.272	0.141	0.144	91.4
	10	0.660	0.576	-0.274	0.139	0.145	90.3
	15	0.753	0.675	-0.279	0.139	0.143	90.6
30	3	0.369	0.296	-0.289	0.122	0.123	92.3
	5	0.490	0.406	-0.286	0.118	0.116	92.6
	7	0.573	0.487	-0.278	0.115	0.119	89.7
	10	0.659	0.575	-0.275	0.115	0.119	88.8
	15	0.752	0.673	-0.283	0.115	0.119	90.5

cluster-level analysis when the first stage model is correctly specified. Again, the results for $\rho = 0.05$ are not presented as we observed qualitatively similar results under $\rho = 0.1$ and 0.05.

Table 8.9 shows the simulation results for the adjusted cluster-level analysis considering no censoring when the first stage model was misspecified. The empirical estimates of RaR were very close to the true value of RaR with good coverage rates which supports our analytical results in Section 8.2.2.

Table 8.4: Simulation results for the unadjusted cluster-level analysis considering only random censoring. Average estimates of log(RaR), their average estimated standard errors (aveSE), empirical standard errors (empSE), and coverage (Cov) rates for nominal 95% confidence interval over 1000 simulation runs are presented. The true log(RaR) is -0.35.

k	Proportions of event		Estimate	aveSE	empSE	Cov (%)
	Control	Intervention				
10	0.022	0.016	-0.348	0.420	0.446	94.7
20			-0.357	0.297	0.310	95.2
30			-0.351	0.245	0.254	95.1
10	0.135	0.100	-0.333	0.261	0.279	94.3
20			-0.335	0.188	0.187	94.7
30			-0.332	0.155	0.159	95.3
10	0.486	0.412	-0.310	0.225	0.237	93.9
20			-0.323	0.163	0.173	93.0
30			-0.311	0.134	0.141	93.5
10	0.696	0.628	-0.310	0.226	0.235	93.6
20			-0.305	0.165	0.163	94.2
30			-0.311	0.135	0.132	94.8

Table 8.5: Simulation results for the adjusted cluster-level analysis considering no censoring when the first stage model is correctly specified. Average estimates of log(RaR), their average estimated standard errors (aveSE), empirical standard errors (empSE), and coverage (Cov) rates for nominal 95% confidence interval over 1000 simulation runs are presented. The true log(RaR) is -0.35.

k	Estimate	aveSE	empSE	Cov (%)
5	-0.352	0.291	0.306	94.6
10	-0.351	0.211	0.215	95.9
20	-0.348	0.150	0.152	93.4
30	-0.354	0.123	0.126	94.6

Table 8.6: Simulation results for the adjusted cluster-level analysis considering only administrative censoring with low to moderate proportions of event when the first stage model is correctly specified. Average estimates of log(RaR), their average estimated standard errors (aveSE), empirical standard errors (empSE), and coverage (Cov) rates for nominal 95% confidence interval over 1000 simulation runs are presented. The true log(RaR) is -0.35.

k	τ	Proportions of event		Estimate	aveSE	empSE	Cov (%)
		Control	Intervention				
5	3	0.051	0.037	-0.358	0.420	0.458	94.8
	5	0.080	0.059	-0.347	0.370	0.406	93.9
	7	0.107	0.079	-0.343	0.348	0.362	94.5
	10	0.145	0.109	-0.348	0.329	0.347	94.4
	15	0.199	0.155	-0.346	0.315	0.334	94.4
10	3	0.051	0.037	-0.351	0.301	0.311	94.5
	5	0.081	0.059	-0.359	0.272	0.284	93.9
	7	0.108	0.080	-0.349	0.253	0.265	94.0
	10	0.144	0.109	-0.345	0.241	0.252	93.7
	15	0.202	0.153	-0.352	0.230	0.248	93.7
20	3	0.051	0.036	-0.351	0.216	0.219	94.5
	5	0.080	0.058	-0.349	0.192	0.190	95.3
	7	0.108	0.080	-0.345	0.181	0.188	95.0
	10	0.145	0.109	-0.357	0.173	0.172	95.1
	15	0.199	0.153	-0.344	0.165	0.165	95.3
30	3	0.051	0.036	-0.348	0.176	0.172	95.2
	5	0.081	0.059	-0.355	0.157	0.161	95.0
	7	0.108	0.080	-0.350	0.148	0.149	94.8
	10	0.146	0.108	-0.352	0.141	0.146	94.0
	15	0.201	0.152	-0.352	0.135	0.138	94.5

8.3 Individual-level analysis

In individual-level analysis, a regression model is fitted to the individual-level outcome, allowing for the fact that observations within the same clusters are correlated. For time-to-event data in CRTs, *shared frailty models* (SFM) are widely used as individual-level analysis method. We now describe SFM briefly.

Table 8.7: Simulation results for the adjusted cluster-level analysis considering administrative censoring with moderate to high proportions of event when the first stage model is correctly specified. Average estimates of log(RaR), their average estimated standard errors (aveSE), empirical standard errors (empSE), and coverage (Cov) rates for nominal 95% confidence interval over 1000 simulation runs are presented. The true log(RaR) is -0.35.

k	τ	Proportions of event		Estimate	aveSE	empSE	Cov (%)
		Control	Intervention				
5	3	0.371	0.296	-0.349	0.301	0.340	93.8
	5	0.492	0.409	-0.340	0.291	0.316	95.9
	7	0.575	0.488	-0.343	0.291	0.300	95.1
	10	0.661	0.576	-0.354	0.284	0.302	94.7
	15	0.750	0.670	-0.348	0.289	0.311	94.4
10	3	0.369	0.295	-0.348	0.218	0.222	94.8
	5	0.489	0.407	-0.342	0.213	0.217	94.9
	7	0.574	0.488	-0.359	0.212	0.217	95.7
	10	0.661	0.575	-0.351	0.212	0.211	94.6
	15	0.751	0.674	-0.347	0.218	0.229	93.4
20	3	0.368	0.296	-0.341	0.155	0.153	95.6
	5	0.491	0.406	-0.344	0.153	0.152	94.6
	7	0.573	0.488	-0.347	0.152	0.152	95.3
	10	0.660	0.576	-0.345	0.151	0.155	95.1
	15	0.753	0.675	-0.352	0.150	0.153	93.9
30	3	0.369	0.296	-0.347	0.128	0.130	95.3
	5	0.490	0.406	-0.348	0.125	0.128	94.0
	7	0.573	0.487	-0.351	0.124	0.128	94.4
	10	0.659	0.575	-0.347	0.124	0.129	93.5
	15	0.752	0.673	-0.351	0.123	0.123	94.9

Table 8.8: Simulation results for the adjusted cluster-level analysis considering only random censoring when the first stage model is correctly specified. Average estimates of log(RaR), their average estimated standard errors (aveSE), empirical standard errors (empSE), and coverage (Cov) rates for nominal 95% confidence interval over 1000 simulation runs are presented. The true log(RaR) is -0.35.

k	Proportions of event		Estimate	aveSE	empSE	Cov (%)
	Control	Interven.				
10	0.281	0.222	-0.353	0.222	0.225	93.9
20			-0.348	0.161	0.161	94.5
30			-0.346	0.131	0.133	94.4
10	0.486	0.411	-0.347	0.215	0.233	93.7
20			-0.354	0.153	0.154	94.6
30			-0.349	0.126	0.127	95.1
10	0.696	0.627	-0.351	0.212	0.220	94.5
20			-0.345	0.153	0.151	94.9
30			-0.345	0.124	0.129	93.8

Table 8.9: Simulation results for adjusted cluster-level analysis considering no censoring when the first stage model is misspecified. Average estimates of log(RaR), their average estimated standard errors (aveSE), empirical standard errors (empSE), and coverage (Cov) rates for nominal 95% confidence interval over 1000 simulation runs are presented. The true log(RaR) is -0.35.

k	Estimate	aveSE	empSE	Cov (%)
5	-0.350	0.287	0.313	94.6
10	-0.347	0.211	0.218	94.8
20	-0.353	0.151	0.155	95.4
30	-0.349	0.124	0.124	94.3

8.3.1 The Shared frailty model

The frailty model is an extension of the Cox model that allows dependency among observations within the same cluster. In a SFM, a multiplicative random effect is common to all individuals of a cluster. The SFM is defined in terms of conditional rate, often referred to conditional hazard, as

$$\lambda_{ijl}(t|X_{ijl}, \delta_{ij}) = \lambda_0(t)\ \delta_{ij} \exp(\beta_1 i + f_i(X_{ijl})), \tag{8.22}$$

where $\lambda_0(t)$ is the baseline rate, $\delta_{ij}(> 0)$ is the frailty for the (ij)th cluster, β_1 is the intervention effect. The conditional rate for the (ijl)th individual is composed of the baseline rate $\lambda_0(t)$, the frailty multiplier δ_{ij} shared by all individuals in the (ij)th cluster, and the adjustment for the covariate X_{ijl}. If the baseline rate $\lambda_0(t)$ is a constant over time, one can write the model (8.22) as

$$\lambda_{ijl}(t|X_{ijl}, \delta_{ij}) = \delta_{ij} \exp(\beta_0 + \beta_1 i + f_i(X_{ijl})) \text{ with } \exp(\beta_0) = \lambda_0(t). \tag{8.23}$$

The main assumption of a SFM is that all individuals in the same cluster share the same frailty value, giving rise to the name. Sharing the same frailty value by all individuals in a cluster generates dependence between event times of two individuals in the same cluster. Conditional on the frailty δ_{ij}, the event times in the (ij)th cluster are assumed to be independent. It is also assumed that event times between clusters are independent. If $\delta_{ij} > 1$, all individuals of the (ij)th cluster are said to have an increased risk of the event. Conversely, if $0 < \delta_{ij} < 1$, all individuals in the (ij)th clusters are less frail and will tend to survive longer period provided that all else is unchanged.

The frailties $\delta_{ij}(i = 0, 1; j = 1, 2, \ldots, k)$ are assumed to be independently and identically distributed random variables with probability density function $f(\delta)$, the frailty distribution. Various frailty distributions have been proposed in the literature [39, 40] including gamma, log-normal, inverse-Gaussian and positive stable distributions. In this thesis, we restrict our attention to the gamma distribution, a family of positively skewed distributions. There are two main reasons for choosing this distribution as the frailty distribution. First, the rates are positive quantities and often have a positively skewed distribution. Second, a simple analytical form for the distribution of the number of events can be derived if the frailties (random effects) follow the gamma distribution. It can be shown that the combination of Poisson distribution with gamma frailties implies that the number of events in clusters follow the negative binomial distribution. Under the assumption that the frailties follow a gamma distribution, the model (8.23) is called gamma SFM. Since the frailties multiply the rate, they need to be non-negative. In addition, in order to separate the baseline rate from the overall effect of random frailties, the mean of frailties is typically constrained to unity. The variance of frailties represents the degree of heterogeneity across the clusters in baseline rate.

8.3.2 Simulation study II

Maximum likelihood methods are used to fit the gamma SFM which are valid asymptotically. However, like LMM and RELR, the SFM could underestimate the standard errors of the parameters estimates when each intervention group has small number of clusters. A simulation study was conducted to investigate the performance of the SFM

for estimating RaR with small number of clusters in each intervention group. We also investigated whether the CIs calculated using quantiles form $t-$distribution give better coverage than that of CIs calculated using quantiles from standard normal distribution.

Data generation and analysis: Data were generated in exactly the same way that we explained in Section 8.2.3. However, in this simulation study, we considered two different cases depending on whether the censoring mechanism is the same or different between the intervention groups. These are (C1) the two intervention groups have the same censoring mechanism, and (C2) the intervention groups have the different censoring mechanism. We set $(\psi_0, \psi_1) = (-1.5, -0.05)$ to generate censoring times under C1, and we set $(\psi_0, \psi_1) = (-1.5, -0.05)$ in the control group and $(\psi_0, \psi_1) = (-2.5, -0.05)$ in the intervention group to generating censoring times under C2. The R package **parfm** [41] was used to fit the gamma SFM (8.23). We calculated the CIs based on quantiles form $\mathcal{N}(0, 1)$ and based on quantiles form $t-$distribution with DF $2k - 2$.

Simulation results: Table 8.10 and Table 8.11 present the average estimates of log(RaR), their average estimated standard errors (aveSE) and empirical standard errors (empSE), and coverage rates for 95% CI over 1000 simulations runs under C1 and C2, respectively, with $\rho = 0.1$. The estimates were unbiased for the true log(RaR) regardless of whether the two intervention groups have the same or different censoring mechanism. However, the average standard error estimates were slightly lower compared to the empirical SEs when the intervention groups had small number of clusters,

Table 8.10: Average estimates of log(RaR) using gamma SFM, their average estimated standard errors (aveSE), empirical standard errors (empSE) and coverage (Cov) rates for nominal 95% CI over 1000 simulations when the two intervention groups have the same censoring mechanism. The true log(RaR) is -0.35. The proportions of event in the control and intervention groups were, respectively, (a) 0.065 and 0.045 (b) 0.407 and 0.333, and (c) 0.855 and 0.795.

	k	Estimate	aveSE	empSE	Cov (%).	
					Normal	$t-$dist.
(a)	5	-0.363	0.369	0.436	89.0	91.0
	10	-0.341	0.273	0.282	93.1	94.7
	20	-0.359	0.198	0.206	93.6	94.8
	50	-0.351	0.127	0.133	93.4	94.3
(b)	5	-0.352	0.274	0.332	86.0	87.3
	10	-0.338	0.210	0.219	93.0	94.8
	20	-0.351	0.154	0.160	93.4	95.1
	50	-0.351	0.099	0.101	94.7	95.6
(c)	5	-0.352	0.266	0.320	87.3	88.9
	10	-0.337	0.203	0.212	92.1	94.2
	20	-0.350	0.149	0.155	92.8	94.5
	50	-0.351	0.096	0.097	94.8	96.1

which resulted in low coverage rate. However, the CIs calculated using quantiles from $t-$distribution showed better coverage rates compared to that of CIs calculated using quantiles from $\mathcal{N}(0,1)$.

8.4 Kaplan-Meier estimator

In this section, the estimand of interest is the survival function in each intervention group and we use the Kaplan-Meier (KM) estimator. Since KM estimator has the same form in each group, we describe it for one intervention group. Therefore, the subscript i is dropped in the remainder of this section.

Table 8.11: Average estimates of log(RaR) using gamma SFM, their average estimated standard errors (aveSE), empirical standard errors (empSE) and coverage (Cov) rates for nominal 95% CI over 1000 simulations when the two intervention groups have different censoring mechanism. The true log(RaR) is -0.35. The proportions of event in the control and intervention groups were, respectively, (a) 0.075 ans 0.054 (b) 0.410 and 0.375, and (c) 0.880 and 0.875

	k	Estimate	aveSE	empSE	Cov (%)	
					Normal	$t-$dist.
(a)	5	-0.358	0.356	0.419	87.7	89.9
	10	-0.345	0.264	0.275	92.6	94.2
	20	-0.358	0.192	0.202	94.2	95.4
	50	-0.352	0.124	0.128	93.4	95.3
(b)	5	-0.352	0.272	0.329	86.6	87.9
	10	-0.348	0.209	0.218	93.4	94.3
	20	-0.351	0.154	0.160	93.6	95.2
	50	-0.351	0.099	0.101	94.4	95.4
(c)	5	-0.353	0.266	0.320	87.4	88.9
	10	-0.342	0.203	0.212	92.3	94.4
	20	-0.350	0.149	0.154	92.9	94.1
	50	-0.351	0.096	0.097	94.9	96.3

Assuming that there are no tied event times, let $t_1 < t_2 < \ldots < t_M$ be the ordered event times, where $M = km$. Define the indicator variables

$$A_{jl}(t_v) = \begin{cases} 1 & \text{if } l\text{th individual from } j\text{th cluster fails at time } t_v \\ 0 & \text{otherwise} \end{cases}$$

and

$$B_{jl}(t_v) = \begin{cases} 1 & \text{if } l\text{th individual from } j\text{th cluster is at risk at time } t_v \\ 0 & \text{otherwise.} \end{cases}$$

The number of events at time t_v is then calculated as

$$d_v = \sum_{j=1}^{k} \sum_{l=1}^{m} A_{jl}(t_v)$$

and the number of individuals at risk at time t_v is

$$n_v = \sum_{j=1}^{k} \sum_{l=1}^{m} B_{jl}(t_v)$$

Ignoring clustering, the KM estimate of survival function at time t_u is

$$\hat{S}(t_u) = \prod_{v=1}^{u} \left(1 - \frac{d_v}{n_v}\right) = \prod_{v=1}^{u} (1 - q_v) = \prod_{v=1}^{u} p_v$$

where $q_v = d_v/n_v$ and $p_v = 1 - q_v$.

Greenwood's formula estimates the variance of the KM estimates, assuming the observations are independent, as

$$\widehat{\text{Var}}\big(\hat{S}(t_u)\big) = \Big\{\hat{S}(t_u)\Big\}^2 \sum_{v=1}^{u} \frac{d_v}{n_v(n_v - d_v)}.$$

However, if the observations are clustered, i.e, the observations in the same cluster are correlated, the variance of KM estimates using Greenwood's formula might be underestimated. Williams (1995) [38] derived a variance estimator for KM estimates which allows for the dependence caused by clustering. This estimator uses a Taylor series linearised approach and the between-cluster variance estimator. We now describe the Williams approach briefly (see the original paper [38] for more details).

Williams' approach used Woodruff's technique [42] that replaces a complex non-linear function, like $\hat{S}(t_v)$, with a linear approximation based on a first-order Taylor series expansion. The linear approximation is then used to estimate the variance of the original non-linear function. Returning to the KM estimate of survival function, $\hat{S}(t_u)$ is a

product of terms containing the ratio $q_v = d_v/n_v, v = 1, 2, \ldots, u$. The linearised value of q_v for the (jl)th individual, using Woodruff's approach [42], is

$$L_{jl}(q_v) = \frac{1}{n_v}\big[A_{jl}(t_v) - q_v B_{jl}(t_v)\big].$$

Then the linearised value for the survival estimate $\hat{S}(t_u)$, developed by Folsom *et.al* [43], is

$$\begin{aligned} L_{jl}[\hat{S}(t_u)] &= -\sum_{v=1}^{u} \frac{\hat{S}(t_u)}{p_v} L_{jl}(q_v) \\ &= -\hat{S}(t_u)\left[\sum_{v=1}^{u} \frac{A_{jl}(t_v) - q_v B_{jl}(t_v)}{n_v - d_v}\right]. \end{aligned}$$

The leading minus sign can be ignored as this will not affect the variance estimate. The linearised values can be calculated using the recursive formula

$$L_{jl}[\hat{S}(t_u)] = p_u L_{jl}\big[\hat{S}(t_{(u-1)})\big] + \hat{S}(t_{u-1}) L_{jl}(q_u);\ u = 2, 3, \ldots, M$$

with $L_{jl}[\hat{S}(t_1)] = L_{jl}(q_1)$. In order to estimate the variance of $\hat{S}(t_u)$, it is assumed that the k clusters are randomly selected from a infinite population of clusters and the individuals within each cluster are correlated. The between-cluster variance estimator can then be applied to the linearised values $L_{jl}[\hat{S}(t_u)]$ to estimate the variance of $\hat{S}(t_u)$.

Accumulating the linearised values to the cluster level as

$$L_j[\hat{S}(t_u)] = \sum_{l=1}^{m} L_{jl}\big[\hat{S}(t_u)\big],$$

the variance of $\hat{S}(t_u)$ is then estimated by

$$\widehat{\text{Var}}[\hat{S}(t_u)] = \frac{k}{k-1}\sum_{j=1}^{k}\left(L_j\left[\hat{S}(t_u)\right] - \bar{L}\left[\hat{S}(t_u)\right]\right)^2,$$

where $\bar{L}(\hat{S}(t_u))$ is the mean of $L_j[\hat{S}(t_u)]$ over j. This variance estimator is unbiased for any linear statistic, and is consistent for non-linear statistic when the number of clusters tends to infinity. Also this method is valid for any correlation structure among the observations within a cluster as long as the clusters are independent. The main advantage of this method is that it does not require any information regarding the within-cluster correlation structure.

In order to construct a confidence interval (CI) for $\hat{S}(t)$ we need to make a distributional assumption. Let $z_{\alpha/2}$ be such that $P\left(Z > z_{\alpha/2}\right) = \alpha/2$, where $Z \sim (0,1)$. Then assuming $\hat{S}(t)$ is normally distributed, an approximate $100(1-\alpha)\%$ confidence interval for $S(t)$ is given by

$$\hat{S}(t) \pm z_{\alpha/2} \times \sqrt{\widehat{\text{Var}}[\hat{S}(t)]}.$$

A drawback of this CI is that the distribution of $\hat{S}(t)$ is not really normal. One possible solution is to transform $\hat{S}(t)$ onto $(-\infty, \infty)$ scale. Consider a complementary log-log transformation

$$V = \log\{-\log[\hat{S}(t)]\}.$$

Applying the delta method, we know

$$\text{Var}[f(U)] \approx \text{Var}(U)\left\{f'[\text{E}(U)]\right\}^2,$$

where $f(U)$ is a function of U. Applying this variance lemma when $U = \hat{S}(t)$, we get

$$\text{Var}\big(-\log[\hat{S}(t)]\big) \approx \frac{\text{Var}[\hat{S}(t)]}{[\hat{S}(t)]^2}.$$

Now, applying the variance lemma, but this time letting $U = -\log[\hat{S}(t)]$, we get

$$\begin{aligned}\text{Var}\big(\log\{-\log[\hat{S}(t)]\}\big) &\approx \frac{\text{Var}\big(-\log[\hat{S}(t)]\big)}{\{-\log[\hat{S}(t)]\}^2} \\ &= \frac{\text{Var}[\hat{S}(t)]}{[\hat{S}(t)]^2\{\log[\hat{S}(t)]\}^2}.\end{aligned}$$

Assuming V is normally distributed, an approximate $100(1-\alpha)\%$ CI for $\log\{-\log[S(t)]\}$ is given by

$$\log\{-\log[\hat{S}(t)\} \pm z_{\alpha/2} \times \sqrt{\text{Var}[\log\{-\log[\hat{S}(t)]\}]},$$

One can then easily obtain a $100(1-\alpha)\%$ CI for $\hat{S}(t)$ by back-transforming as

$$\left([\hat{S}(t)]^{\exp\{z_{\alpha/2}\sqrt{\text{Var}[\log\{-\log[\hat{S}(t)]\}]}\}}, [\hat{S}(t)]^{\exp\{-z_{\alpha/2}\sqrt{\text{Var}[\log\{-\log[\hat{S}(t)]\}]}\}}\right).$$

As far as we are aware no study has been done to investigate the performance of Williams [38] approach for estimating the SEs of KM estimates in the CRT setup.

8.4.1 Simulation study III

A simulation study was conducted to investigate the performance of Greenwood and Williams approaches for estimating SEs of KM estimates when the interventions groups have low, moderate or high proportions of event and the value of ICC (ρ) is small. Williams [38] conducted a similar simulation study with correlated time-to-event data

but he considered high values for ρ= (0.1,0.3,0.5). However, in practice, the value of ρ in CRTs typically ranges from 0.001 to 0.05 in primary care and health research, and it is rare to have ICC above 0.1 [3]. It is not clear from Williams' paper [38] how his approach performs compare to Greenwood approach when the value of ρ is small and the intervention groups have low, moderate or high proportions of event.

Data generation and analysis: For each individual in the study, the baseline covariate X_{ijl} and event time T_{ijl} were generated exactly the same way that we explained in Section 8.2.3. We set $\rho = (0.1, 0.05, 0.001)$. The independent censoring times were generated as $C_{ijl} \sim \text{Exp}(0.5)$. Then we observed the event time $Y_{ijl} = \min(T_{ijl}, C_{ijl}, \tau = 3)$ and the event indicator $\Delta_{ijl} = 1$ if $T_{ijl} < C_{ijl}$ and $T_{ijl} < 3$, and 0 otherwise.

We calculated the KM estimates from a very large data set with 500 clusters in each intervention group and 500 individuals in each cluster. The estimated survival probabilities at six different time-points (0.5,1.0,1.5,2.0,2.5,3.0) were calculated and used as the true probabilities for these time points. In the simulation study, we fixed the number of clusters in each intervention group as $k = 20$ and the cluster size as $m = 100$. We considered (a) small, (b) moderate and (c) high proportions of event in the interventions groups by varying the value of β_0 in equation (8.21). Then for each generated dataset, we calculated the KM estimates at these six points, their standard errors using Greenwood's and Williams' approach, and 95% CI.

Table 8.12: Average KM estimates at the selected time-points in the control group, their empirical SE (empSE) and average estimated SE (aveSE) using Greenwood's and Williams' approaches, and corresponding coverage rates for nominal 95% CI over 1000 simulation runs. The proportions of events in the control and intervention groups were, respectively, (a) 0.067 and 0.035 (b) 0.299 and 0.190, and (c) 0.843 and 0.747. The intra-class correlation coefficient was 0.1.

	Time-points	Survival prob.		empSE	aveSE		Coverage rate (%)	
		True	Average Estimate		Green-wood	Williams	Green-wood	Williams
(a)	0.5	0.976	0.976	0.005	0.004	0.004	87.6	92.0
	1.0	0.954	0.954	0.007	0.005	0.007	85.4	93.2
	1.5	0.932	0.933	0.010	0.007	0.010	81.3	93.3
	2.0	0.912	0.913	0.012	0.008	0.012	80.1	92.4
	2.5	0.894	0.894	0.015	0.010	0.014	78.6	92.3
	3.0	0.876	0.877	0.017	0.011	0.016	80.9	92.9
(b)	0.5	0.852	0.854	0.016	0.008	0.016	66.4	93.9
	1.0	0.748	0.751	0.024	0.011	0.023	62.2	93.0
	1.5	0.670	0.672	0.028	0.013	0.028	62.3	93.5
	2.0	0.606	0.608	0.031	0.014	0.030	62.0	93.1
	2.5	0.555	0.556	0.033	0.016	0.032	63.6	93.1
	3.0	0.510	0.511	0.035	0.017	0.034	66.3	93.0
(c)	0.5	0.226	0.227	0.028	0.010	0.027	53.7	91.5
	1.0	0.111	0.111	0.020	0.008	0.019	60.1	90.6
	1.5	0.066	0.067	0.015	0.007	0.014	67.1	91.5
	2.0	0.044	0.044	0.012	0.006	0.011	71.4	91.0
	2.5	0.031	0.031	0.010	0.006	0.009	76.3	90.9
	3.0	0.023	0.023	0.009	0.005	0.008	81.6	90.8

Simulation results: The average KM estimates at the six considered time-points of the survival function of the control group, their empirical and average estimated SEs using Greenwood and Williams approaches, and their corresponding coverage rates over 1000 simulation runs are presented in Table 8.12, Table 8.13 and Table 8.14, respectively, for $\rho = (0.1, 0.05, 0.01)$. We considered three scenarios for proportion of event: (a) low, (b) moderate and (iii) high. As a result, the true survival probabilities at a particular time-point across the scenarios were different. As expected, the average KM estimates

Table 8.13: Average KM estimates at the selected time-points in the control group, their empirical SE (empSE) and average estimated SE (aveSE) using Greenwood's and Williams' approaches, and corresponding coverage rates for nominal 95% CI over 1000 simulation runs. The proportions of events in the control and intervention groups were, respectively, (a) 0.027 and 0.019 (b) 0.305 and 0.246, (c) 0.710 and 0.643. The true intraclass correlation coefficient was 0.05.

	Time-points	Survival prob.		empSE	aveSE		Coverage rate (%)	
		True	Average Estimate		Green-wood	Williams	Green-wood	Williams
(a)	0.5	0.991	0.991	0.002	0.002	0.002	94.4	94.2
	1.0	0.982	0.982	0.004	0.003	0.004	93.1	94.7
	1.5	0.974	0.974	0.005	0.004	0.004	92.8	94.8
	2.0	0.966	0.965	0.006	0.005	0.006	91.4	93.6
	2.5	0.957	0.957	0.007	0.006	0.007	91.5	94.7
	3.0	0.949	0.949	0.008	0.007	0.008	91.3	94.1
(b)	0.5	0.849	0.852	0.014	0.008	0.014	76.2	91.7
	1.0	0.743	0.747	0.020	0.011	0.020	73.0	93.7
	1.5	0.661	0.666	0.023	0.013	0.023	70.8	94.2
	2.0	0.597	0.601	0.025	0.014	0.025	72.0	93.5
	2.5	0.542	0.546	0.027	0.016	0.027	74.3	94.8
	3.0	0.495	0.500	0.028	0.017	0.028	75.5	94.2
(c)	0.5	0.442	0.446	0.025	0.012	0.025	64.5	94.5
	1.0	0.274	0.275	0.023	0.011	0.023	68.5	94.1
	1.5	0.190	0.190	0.020	0.011	0.020	72.2	93.9
	2.0	0.140	0.140	0.018	0.010	0.018	77.3	94.2
	2.5	0.107	0.107	0.016	0.010	0.016	80.0	94.7
	3.0	0.084	0.085	0.014	0.010	0.014	83.0	92.8

were very close to the true values at all considered six points regardless of the values of ρ and the proportions of event. The Greenwood SEs estimates were lower than the empirical SEs and, consequently, the coverage rates were lower compared to the nominal rate unless the proportion of event was low and the value of ρ was small (0.05,0.01). In contrast, the SEs estimates using Williams approach were very close to the empirical SEs regardless of the proportions of event and the value of ρ and, consequently, the coverage rates were also very close to the nominal rate. Both approaches performed

Table 8.14: Average KM estimates at the selected time-points in the control group, their empirical SE (empSE) and average estimated SE (aveSE) using Greenwood's and Williams' approaches, and corresponding coverage rates for nominal 95% CI over 1000 simulation runs. The proportions of events in the control and intervention groups were, respectively, (a) 0.027 and 0.019 (b) 0.311 and 0.249, (c) 0.720 and 0.654. The true intraclass correlation coefficient was 0.01.

	Time-points	Survival prob.		empSE	aveSE		Coverage rate (%)	
		True	Average Estimate		Green-wood	Williams	Green-wood	Williams
(a)	0.5	0.991	0.991	0.002	0.002	0.002	94.9	93.9
	1.0	0.982	0.982	0.004	0.003	0.003	93.6	93.5
	1.5	0.973	0.973	0.005	0.004	0.005	94.1	93.8
	2.0	0.965	0.965	0.006	0.005	0.006	93.5	94.3
	2.5	0.957	0.957	0.007	0.006	0.007	93.2	93.6
	3.0	0.949	0.949	0.008	0.007	0.008	93.4	93.9
(b)	0.5	0.851	0.851	0.012	0.008	0.011	84.3	92.3
	1.0	0.745	0.743	0.016	0.011	0.016	82.6	93.8
	1.5	0.662	0.659	0.019	0.013	0.018	81.9	93.2
	2.0	0.593	0.592	0.021	0.014	0.020	81.8	92.5
	2.5	0.539	0.537	0.022	0.016	0.021	83.3	92.0
	3.0	0.491	0.489	0.023	0.017	0.023	85.1	93.2
(c)	0.5	0.433	0.435	0.018	0.012	0.019	78.5	95.1
	1.0	0.261	0.262	0.017	0.011	0.017	80.7	94.5
	1.5	0.176	0.177	0.015	0.011	0.015	83.9	93.8
	2.0	0.127	0.128	0.013	0.010	0.013	86.4	93.8
	2.5	0.096	0.097	0.012	0.010	0.012	88.6	94.3
	3.0	0.075	0.076	0.011	0.009	0.011	90.0	93.9

similarly for estimating SEs when the proportion of event was low and the value of ρ was small. We observed qualitatively similar results for the intervention group and the results are not presented in this thesis. One can then compare the KM estimates of the survival curves in the control and intervention groups at a particular point of follow-up period using standard $t-$test. The R code for estimating SEs of KM estimates using Williams [38] approach are given in **Appendix B**.

8.5 Summary

In this chapter, first, we investigated under which conditions the cluster-level analysis methods for analysing time-to-event outcomes in CRTs are consistent. In the case of no censored observations in the data, we showed that the unadjusted cluster-level analysis for estimating RaR is consistent when the covariate effects are the same between the intervention groups. With censored data, the unadjusted cluster-level analysis is consistent when the event rates are small between the intervention groups and the covariate effects are the same between the intervention groups. In contrast, the adjusted cluster-level estimator for RaR is consistent regardless of whether there are censored observations or not when the covariate effects are the same between the intervention groups. However, in the case of censored data, the first stage model needs to be correctly specified.

Second, we investigated the performance of gamma SFM as individual-level analysis when the number of clusters is small in each intervention group. We found that it underestimated the SEs of the RaR estimates when each intervention group have small number of clusters and, consequently, resulted in low coverage. In this case, we also found that the CIs calculated using $t-$distribution gave better coverage than that of CIs calculated using standard normal distribution.

Finally, we compared the performance of Greenwood and Williams approaches for estimating the SEs of KM estimates of survival function in the setting of CRTs with small ICC. We found that the SEs estimates of KM estimates using Williams approach are very close to empirical standard errors and, consequently, the coverage rates are very

close to the nominal rate. In contrast, Greenwood approach underestimated the standard errors of KM estimates and, consequently, resulted in low coverage rate unless the event rate is small and the value of ICC is small.

In practice, it is common to have censored data almost always. We recommend to use unadjusted cluster-level analysis when the event rates are small between the intervention groups, if one is willing to assume that the covariate effects are the same between the intervention groups. In the case of adjusting for baseline covariates in cluster-level analysis, adjusted cluster-level analysis can be used when the analyst can correctly models the dependence on covariates, if the covariate effects are the same between the intervention groups. In case of individual-level analysis with large number of clusters, the gamma SFM can be used if one is willing to assume constant rate over time.

Part V

Discussion and Conclusions

Chapter 9

Discussion and Conclusion

The aim of this thesis was to investigate the validity of methods for the analysis of CRTs for the three outcome types: continuous, binary and time-to-event, when outcomes are missing under the assumption of CDM mechanism. In this final chapter, we review our work presented in this thesis, highlighting the key findings, and outlining possible areas of interest for future research. We summarise our findings for continuous, binary and time-to-event outcomes in Section 9.1, Section 9.2 and Section 9.3, respectively. We outline areas of interest for future research in Section 9.4. We give some concluding remarks in Section 9.5.

9.1 Continuous outcomes

In Part II, we considered continuous outcomes. We investigated the impact of cluster mean imputation for missing outcomes, under MCAR and MAR, on the validity of the ANOVA estimators for the variance components, namely, within-cluster variance and between-cluster variance. We also investigated the impact of CRA and cluster mean imputation for missing outcomes on the validity and power of cluster-level $t-$test, adjusted $t-$test and LMM under MCAR and MAR. Then we investigated the performance of cluster-level analyses and LMM when outcomes are missing under the assumption of CDM mechanism.

Cluster mean imputation has been recommended as a valid approach for handling missing outcomes [17]. In Chapter 4, we showed that the ANOVA estimators of the variance components are biased with cluster mean imputation. The estimate of ICC is also biased. Therefore, we do not recommend cluster mean imputation, since the variance components and ICC are often of interest in CRTs. We also showed that cluster-level $t-$test, adjusted $t-$test and LMM give similar power with full data, CRA and cluster mean imputation, when cluster sizes do not vary largely. In this situation, the cluster-level $t-$test could be an attractive option for testing intervention effect because of its simplicity compared to both adjusted $t-$test and LMM. However, when observed cluster sizes vary to a greater extent, adjusted $t-$test and LMM give better power compared to cluster-level $t-$test using CRA at small values of ICC.

In Chapter 5 (research paper I), we showed that cluster-level analyses are in general biased using CRA unless the two intervention groups have the same missingness mechanism and the same covariate effects on outcome in the data generating model, which is arguably unlikely to hold in practice. We therefore caution researchers that these methods may commonly give biased inferences in CRTs when outcomes are missing under CDM mechanism. In the case of individual-level analysis, we showed that LMM using CRA adjusted for covariates such that the CDM assumption holds gives unbiased estimates of intervention effect regardless of whether missingness mechanisms are the same or different between the intervention groups, and whether there is an interaction between intervention and baseline covariate in the data generating model for the outcome, provided that such interaction is included in the model when required. We also found that there is no gain in terms of bias or efficiency of the estimates using MMI over CRA adjusted for covariates such that the CDM assumption holds as long as both approaches use the same functional form of the same set of baseline covariates and the same modelling assumptions. Therefore, where the CDM assumption for missing outcomes is plausible, and in the absence of auxiliary variables, we recommend that LMM using CRA adjusted for covariates such that the CDM assumption holds as the primary analysis approach for CRTs with missing outcomes.

9.2 Binary outcomes

In Part III, we considered binary outcomes. In this part, first, we derived sufficient conditions for the consistency of the adjusted cluster-level analysis for RR with full data. Then we investigated the validity of RD and RR as measures of intervention effect

using cluster-level analyses when outcomes are missing under the assumption of CDM mechanism. We also investigated the performance of RELR and GEE as individual-level analysis approaches under the same CDM assumption, considering the limitations of previous studies [25–28], which we described in Chapter 6.

In Chapter 7 (research paper II), firstly, we showed that the adjusted cluster-level estimator of RR using full data is consistent and, therefore, asymptotically unbiased for true RR if the true data generating model is a log link model, the functional form of the covariates is the same between the intervention groups, and the distribution of random effect is the same between the intervention groups. Then, we showed that cluster-level analyses for estimating RD using CRA are in general biased under CDM assumption. For estimating RR, both unadjusted and adjusted cluster-level analyses using CRA are valid if the true data generating model has log link and the intervention groups have the same missingness mechanism and the same functional form of the covariates in the outcome model. In contrast, MMI followed by cluster-level analyses gives valid inferences for estimating RD and RR regardless of whether the missingness mechanisms are the same or different between the intervention groups, and whether there is an interaction between intervention and baseline covariate in the outcome model, provided that such interaction is included in the imputation model when required. An alternative often used in the trials context to allow for such an interaction is to impute separately in the two intervention groups.

In the case of individual-level analysis, both RELR and GEE give valid inferences using both CRA adjusted for covariates such that the CDM assumption holds and MMI regardless of whether the missingness mechanisms are the same or different between

the intervention groups, and whether there is an interaction between intervention and baseline covariate in the outcome model, provided that such interaction is included in both the imputation model and the analysis model when required. This conclusion regarding the performance of RELR contradicts the results of a previous study by Ma *et al* [27], where they concluded that RELR using CRA gives biased inference under CDM assumption. We believe our results and explanations help in understanding some of the surprising results and conclusions in Ma *et al* [25–27]. As was the case with continuous outcomes, in the absence of auxiliary variables, there is no benefit in performing MMI rather than doing CRA, under the CDM assumption. Therefore, in the absence of auxiliary variables, and where the CDM assumption for missing outcomes is plausible, we recommend RELR and GEE using CRA adjusted for covariates such that the CDM assumption holds as the primary analysis approach for CRTs with missing outcomes.

9.3 Time-to-Event outcomes

In Part IV, we considered time-to-event outcomes. First, we investigated the consistency of the cluster-level methods for estimating RaR. In the case of no censored observations in the data, we showed that the unadjusted cluster-level analysis for estimating RaR is consistent when the intervention groups have the same covariate effect. In the case of censored data, we showed that the unadjusted cluster-level analysis is consistent when the event rates are small and the covariate effects are the same between the intervention groups. In contrast, the adjusted cluster-level estimator for RaR is consistent regardless

of whether there are censored observations or not when the covariate effects are the same between the intervention groups. However, in the case of censored data, the first-stage model needs to be correctly specified the dependence on covariates.

Second, we investigated the performance of the gamma SFM as an individual-level analysis when the number of cluster is small in each intervention group. We found that it underestimates the SEs of the estimates when each intervention group has small number of clusters.

Finally, we compared the performance of Greenwood and William approaches for estimating the SEs of KM estimates of survival function in the setting of CRTs. We showed that the SEs estimates of KM estimates using Williams approach are very close to empirical SEs and, consequently, the coverage rates are very close to the nominal rate. In contrast, Greenwood approach underestimates the SEs of KM estimates and, consequently, resulted in low coverage rates unless the event rates are small and the value of ICC is small.

Since censored data are common almost always in practice, we can make the following recommendations based on our analytical and simulation results. We recommend to use unadjusted cluster-level analysis when the event rates are small between the intervention groups, if one is willing to assume that the covariate effects are the same between the intervention groups. In the case of adjusting for baseline covariates in cluster-level analysis, adjusted cluster-level analysis can be used when the analyst can correctly models the dependence on covariates, if the covariate effects are the same between the intervention groups. In case of individual-level analysis with large number of clusters,

the gamma shared frailty model can be used if one is willing to assume constant rate over time. However, the gamma shared frailty model underestimates the SEs of the parameters estimates when the number of clusters is small.

9.4 Future work

In this thesis, we assumed baseline CDM mechanism for missing outcomes in CRTs which is an example of MAR when the covariates are fully observed. In practice, given the observed data, we cannot identify which missingness assumption is appropriate [44, 45]. Inferences obtained under the CDM assumption may not be valid if this assumption does not hold. It is therefore imperative to explore the robustness of the inferences under a range of plausible MNAR missingness mechanisms [44]. This is known as sensitivity analysis. The idea is to analyse the data assuming a range of plausible MNAR mechanisms for the missing outcomes and see how robust the inferences are across the different mechanisms. The degree to which inferences are robust across a range of plausible MNAR mechanisms indicates how sensitive conclusions are to missing outcomes to the CDM/MAR assumption.

Analysing partially observed outcomes under a MNAR missingness mechanism is more complex. There are three broad types of modelling approach which can be applied. These approaches are shared parameter modelling, selection modelling and pattern mixture modelling. The shared parameter modelling approach uses a set of latent variables (random effects) to model the relationship between missingness and the outcome [46]. The selection modelling approach specifies a model for how missingness depends both

on the observed and unobserved outcome data. The pattern mixture modelling approach assumes a distribution for the missing outcomes which may be different for each missingness pattern. All these three approaches can be applied with maximum likelihood methods, Bayesian methods and MI [44, 47]. These methods have been developed extensively in the context of non-clustered data. Further research is needed on how to extent and apply these approaches for sensitivity analysis in CRTs. For example, in this thesis, we have explored MMI methods assuming MAR. These methods could be modified to do MNAR sensitivity analysis in the CRT setting.

9.5 Concluding remarks

As we discussed in Chapter 2, missing outcomes are very common in CRTs [11, 12]. Handling such data is one of the main challenges faced by an analyst wishing to analyse a CRT. Although CRTs are increasingly being used to evaluate the effectiveness of interventions in health services research [1, 2], there is limited guidance on how to handle missing data in CRTs. This thesis investigated the validity of the methods for the analysis of CRTs for the three common outcome types when outcomes are missing under CDM mechanism. We gave recommendations based on our analytical and simulations results for which methods to use to get valid inferences despite having missing outcomes under CDM assumption. The choice of appropriate methods depend on type of outcome and parameter of interest. We hope this thesis will help researchers to choose appropriate methods to get valid inferences from CRTs, when it is assumed that outcomes are missing under CDM mechanism.

Bibliography

[1] Donner A, Klar N. Pitfalls of and controversies in cluster randomization trials. American Journal of Public Health. 2004;94(3):416–422.

[2] Donner A, Klar N. Design and Analysis of Cluster Randomization Trials in Health Research. Arnold, London; 2000.

[3] Murray DM, Blitstein JL. Methods to reduce the impact of interclass correlation in group-randomised trials. Evaluation Review. 2003;27:79–103.

[4] Murray DM. Design and Analysis of Group- Randomized Trials. Oxford University Press, New York; 1998.

[5] Hayes RJ, Moulton LH. Cluster Randomised Trials. CRC Press, Taylor & Francis Group; 2009.

[6] Hernandez AV, Steyerberg EW, Habbema JD. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. Journal of Clinical Epidemiology. 2004;57(5):454–460.

[7] Gail MH, Tan WY, Piantadosi S. Tests for no treatment effect in randomised clinical trials. Biometrika. 1988;75(1):57–64.

[8] Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. Clinical Trials. 2004;1(4):368–376.

[9] Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338.

[10] Andridge RR. Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. Biometrical Journal. 2011;53(1):57–74.

[11] Diaz-Ordaz K, Kenward MG, Cohen A, Coleman CL, Eldridge S. Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. Clinical Trials. 2014;11(5):590–600.

[12] Fiero MH, Huang S, Oren E, Bell ML. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. Trials. 2016;17(1):72.

[13] Little RJA, Rubin DB. Statistical Analysis with Missing Data. 2nd Edition. John Wiley & Sons, New Jersey.; 2002.

[14] Rubin DB. Inference and missing data. Biometrika. 1976;63:581–592.

[15] Groenwold RH, Donders AR, Roes KC, Harrell FE, Moons KG. Dealing with missing outcome data in randomized trials and observational studies. American Journal of Epidemiology. 2012;175(3):210–217.

[16] Molenberghs G, Kenward MG. Missing data in clinical studies. John Wiley & Sons, Chichester, UK; 2007.

[17] Taljaard M, Donner A, Klar N. Imputation Strategies for Missing Continuous Outcomes in Cluster Randomized Trials. Biometrical journal. 2008;50(3):329–345.

[18] Rubin DB. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons , New York; 1987.

[19] Buuren SV. Flexible Imputation of Missing Data. CRC press, Taylor & Francis Group; 2012.

[20] Meng XL. Multiple-imputation inferences with uncongenial sources of input. Statistical Science. 1994;9(4):538–558.

[21] DiazOrdaz K, Kenward M, Gomes M, Grieve R. Multiple imputation methods for bivariate outcomes in cluster randomised trials. Statistics in Medicine. 2016;epub.

[22] Gomes M, Diaz-Ordaz K, Grieve R, Kenward MG. Multiple imputation methods for handling missing data in cost-effectiveness analyses that use data from hierarchical studies: An application to cluster randomized trials. Medical Decisions Making. 2013;33(8):1051–1063.

[23] Díaz-Ordaz K, Kenward MG, Grieve R. Handling missing values in cost effectiveness analyses that use data from cluster randomized trials. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2014;177(2):457–474.

[24] Rubin DB, Schenker N. Multiple imputation for interval estimation from simple random samples with ignorable nonresponses. Journal of the American Statistical Association. 1986;81(394):366–374.

[25] Ma J, Akhtar-Danesh N, Dolovich L, Thabane L, Chambers LW, Kaczorowski J, et al. Imputation strategies for missing binary outcomes in cluster randomized trials. BMC Medical Research Methodology. 2011;11:18.

[26] Ma J, Raina P, Beyene J, Thabane L. Comparing the performance of different multiple imputation strategies for missing binary outcomes in cluster randomized trials: a simulation study. J Open Access Med Stat. 2012;2:93–103.

[27] Ma J, Raina P, Beyene J, Thabane L. Comparison of population-averaged and cluster-specific models for the analysis of cluster randomized trials with missing binary outcomes: a simulation study. BMC Medical Research Methodology. 2013;13:9.

[28] Caille A, Leyrat C, Giraudeau B. A comparison of imputation strategies in cluster randomized trials with missing binary outcomes. Statistical Methods in Medical Research. 2014;epub.

[29] Donner A, Klar N. Confidence interval construction for effect measures arising from cluster randomization trials. Journal of Clinical Epidemiology. 1993;46:123 – 131.

[30] Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihoods. Biometrics. 1997;53:983–997.

[31] Verbeke G, Molenberghs G. Linear Mixed Models for Longitudinal Data. Springer.; 2000.

[32] Satterthwaite FE. Synthesis of variance. Psychometrika. 1941;6:390–316.

[33] Faes C, Molenberghs G, Aerts M, Verbeke G, Kenward MG. The effective sample size and an alternative small-sample degrees-of-freedom method. The American Statistician. 2009;63(4):389–399.

[34] Ukoumunne OC, Carlin JB, Gulliford MC. A simulation study of odds ratio estimation for binary outcomes from cluster randomized trials. Statistics in Medicine. 2007;26(18):3415–3428.

[35] Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. American Journal of Public Health. 2004;94(3):423–432.

[36] Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. Biometrics. 2001;57(1):126–134.

[37] Davies HTO, Crombie IK, Tavakoli M. When can odds ratios mislead? BMJ. 1998;316(7136):989–991.

[38] Williams RL. Product-limit survival functions with correlated survival times. Lifetime data analysis. 1995;1(2):171–186.

[39] Duchateau L, Janssen P. The frailty model. Springer Science & Business Media; 2007.

[40] Wienke A. Frailty models in survival analysis. CRC Press; 2010.

[41] Munda M, Rotolo F, Legrand C, et al. parfm: Parametric frailty models in R. Journal of Statistical Software. 2012;51(1):1–20.

[42] Woodruff RS. A simple method for approximating the variance of a complicated estimate. Journal of the American Statistical Association. 1971;66(334):411–414.

[43] Folsom R, Lavange L, Williams RL. A probability sampling perspective on panel data analysis. In Panel Surveys (Kasprzyk D, Duncan G, Kalton G and Singh MP, eds). 1989;p. 108–138.

[44] Carpenter JR, Kenward MG. Multiple Imputations and its Applications. John Wiley & Sons; 2013.

[45] White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Statistics in Medicine. 2010;29(28):2920–2931.

[46] Creemers A, Hens N, Aerts M, Molenberghs G, Verbeke G, Kenward MG. A Sensitivity Analysis for Shared-Parameter Models for Incomplete Longitudinal Outcomes. Biometrical Journal. 2010;52(1):111–125.

[47] Molenberghs G, Fitzmaurice G, Kenward MG, Tsiatis A, Verbeke G. Handbook of missing data methodology. CRC Press; 2014.

Appendices

Appendix A:

Table A1: Empirical type I error rate over 1000 simulation runs of LMM using the $z-$test and the Wald $t-$test (using Satterthwaite's approximation for degrees freedom) for intervention effect with CRA and cluster mean imputation for missing values under MCAR2.

k	m	ρ	Full data		CRA		cluster mean imputation	
			$z-$ test	Wald $t-$test	$z-$ test	Wald $t-$test	$z-$ test	Wald $t-$test
5	30	0.01	4.8	3.6	5.4	4.4	**8.1**	4.9
		0.05	**7.6**	4.2	**7.3**	5.4	**8.5**	5.0
		0.10	**8.3**	5.1	**7.8**	4.3	**7.4**	4.7
	50	0.01	6.8	4.6	5.1	3.9	**7.5**	4.3
		0.05	**9.0**	5.8	**7.8**	4.7	**7.4**	4.4
		0.10	**8.9**	5.2	**7.5**	4.5	**7.4**	4.0
	100	0.01	6.2	4.3	**7.0**	4.6	**8.5**	5.5
		0.05	**7.9**	3.7	**9.0**	5.2	**8.9**	4.5
		0.10	**7.9**	4.1	**8.8**	5.3	**9.2**	5.0
	250	0.01	**8.5**	5.4	**9.1**	6.1	**8.7**	5.4
		0.05	**8.6**	4.6	**9.1**	5.0	**9.1**	5.2
		0.10	**8.6**	4.4	**9.2**	5.1	**9.0**	4.9
10	30	0.01	5.8	5.0	5.3	4.2	6.4	5.2
		0.05	**7.6**	6.4	**7.6**	5.4	6.1	4.7
		0.10	**7.4**	6.0	**7.3**	5.3	6.4	4.9
	50	0.01	5.5	4.8	5.7	4.3	5.7	4.1
		0.05	**7.4**	5.3	6.0	4.2	6.5	4.6
		0.10	**7.3**	5.6	6.1	4.3	6.3	4.2
	100	0.01	5.0	3.7	4.6	3.3	5.2	3.8
		0.05	5.9	4.0	5.2	3.6	4.9	3.6
		0.10	5.9	4.6	4.7	3.6	4.9	3.8
	250	0.01	4.7	3.4	6.2	4.9	5.9	4.3
		0.05	5.6	4.4	6.7	5.1	6.8	5.0
		0.10	5.7	4.6	6.5	4.9	6.4	4.7
15	30	0.01	5.5	4.9	6.5	6.0	7.1	5.9
		0.05	6.1	5.5	5.9	5.2	6.1	5.3
		0.10	5.9	5.3	5.7	4.5	5.9	4.8
	50	0.01	6.4	5.4	5.6	5.1	5.6	4.9
		0.05	5.5	4.9	6.3	5.5	6.0	5.4
		0.10	5.7	4.6	6.1	5.7	6.2	5.3

Table A2: Empirical type I error rate over 1000 simulation runs of LMM using the $z-$test and the Wald $t-$test (using Satterthwaite's approximation for degrees freedom) for intervention effect with CRA and cluster mean imputation for missing values under MAR.

k	m	ρ	Full data		CRA		cluster mean imputation	
			$z-$ test	Wald $t-$test	$z-$ test	Wald $t-$test	$z-$ test	Wald $t-$test
5	30	0.01	4.8	3.6	**6.6**	5.0	**8.6**	5.5
		0.05	**7.6**	4.2	**7.3**	5.1	**8.1**	4.7
		0.10	**8.3**	5.1	**8.2**	5.3	**8.2**	5.0
	50	0.01	6.8	4.6	6	5.1	**8.0**	5.7
		0.05	**9.0**	5.8	**8.1**	5.1	**8.5**	4.8
		0.10	**8.9**	5.2	**8.7**	5.1	**8.7**	5.1
	100	0.01	6.2	4.3	6.2	3.9	**7.5**	4.0
		0.05	**7.9**	3.7	**8.7**	4.7	**8.9**	4.4
		0.10	**7.9**	4.1	**8.4**	4.7	**8.4**	4.9
	250	0.01	**8.5**	5.4	**8.5**	4.9	**8.9**	4.5
		0.05	**8.6**	4.6	**9.6**	4.6	**9.5**	4.6
		0.10	**8.6**	4.4	**9.1**	5.1	**8.9**	5.2
10	30	0.01	5.8	5.0	4.9	4.1	6.1	4.9
		0.05	**7.6**	6.4	6.3	5.1	6.2	4.8
		0.10	**7.4**	6.0	6.5	5.0	6.4	4.6
	50	0.01	5.5	4.8	5.0	4.2	6.6	5.0
		0.05	**7.4**	5.3	6.9	5.6	6.9	5.6
		0.10	**7.3**	5.6	6.7	5.0	6.6	5.1
	100	0.01	5.0	3.7	**7.1**	5.7	**7.5**	5.7
		0.05	5.9	4.0	**7.4**	5.3	**7.6**	5.6
		0.10	5.9	4.6	**7.1**	5.0	**7.2**	5.0
	250	0.01	4.7	3.4	**7.8**	6.6	**8.2**	6.6
		0.05	5.6	4.4	**8.5**	5.7	**8.4**	5.7
		0.10	5.7	4.6	**7.5**	5.5	**7.5**	5.7
15	30	0.01	5.5	4.9	4.4	4.1	5.6	4.4
		0.05	6.1	5.5	5.1	3.7	5.3	4.1
		0.10	5.9	5.3	5.3	4.2	5.2	4.3
	50	0.01	6.4	5.4	4.0	3.4	4.7	3.5
		0.05	5.5	4.9	5.5	4.3	5.7	4.6
		0.10	5.7	4.6	5.7	4.7	5.9	4.8
	100	0.01	5.1	4.1	5.6	4.7	6.0	5.1
		0.05	5.4	4.4	4.9	4.0	4.9	3.9
		0.10	5.6	4.4	5.1	4.1	5.0	4.0

Table A3: Empirical power values of the cluster-level $t-$test, adjusted $t-$test and LMM with Wald $t-$test for intervention effect over 1000 simulation runs using full data, CRA and cluster mean imputation for missing values under MCAR2.

k	m	ρ	Full data		CRA			LMM with cluster mean imputation
			Cluster level $t-$test	LMM approach	Cluster level $t-$test	Adjusted $t-$test	LMM approach	
5	30	0.01	63.8	60.9	49.7	49.6	47.8	46.7
		0.05	38.0	38.2	33.1	32.4	32.3	33.1
		0.10	25.6	26.6	22.9	22.4	22.3	22.9
	50	0.01	76.6	74.8	62.6	64.5	64.2	62.9
		0.05	42.9	39.6	36.7	35.6	36.6	36.7
		0.10	28.3	25.0	23.6	23.0	23.5	23.5
	100	0.01	90.5	89.2	82.6	83.7	83.9	82.6
		0.05	48.7	46.8	45.0	43.7	45.1	45.0
		0.10	28.7	29.3	28.9	29.3	28.9	28.9
	250	0.01	97.3	97.2	95.2	94.6	95.2	95.2
		0.05	54.2	51.7	49.9	49.2	49.4	49.9
		0.10	31.6	33.0	29.7	28.5	29.7	29.7
10	30	0.01	92.0	93.4	78.5	82.4	81.2	78.5
		0.05	68.2	71.0	60.5	61.5	61.6	60.5
		0.10	50.7	51.2	43.9	43.6	43.7	43.9
	50	0.01	98.4	97.7	93.1	94.6	94.5	93.1
		0.05	76.4	77.6	71.3	71.4	71.7	71.3
		0.10	52.1	54.8	52.3	50.4	52.8	52.3
	100	0.01	99.7	99.8	99.5	99.2	99.5	99.5
		0.05	82.7	84.1	78.5	79.7	79.3	78.5
		0.10	57.3	58.9	54.0	53.7	54.2	54.0
20	30	0.01	99.8	100	98.6	98.9	98.8	98.6
		0.05	94.3	96.0	90.5	90.0	90.9	90.5
		0.10	80.0	80.9	76.5	75.3	76.8	76.5
	50	0.01	100	100	100	100	100	100
		0.05	97.0	97.1	94.3	94.4	94.6	94.3
		0.10	84.0	86.1	80.9	79.9	80.4	80.9
	100	0.01	100	100	100	100	100	100
		0.05	98.4	98.9	98.1	98.3	98.3	98.1
		0.10	87.3	87.0	85.8	84.1	86.3	85.8

Table A4: Empirical power values of the cluster-level $t-$test, adjusted $t-$test and LMM with Wald $t-$test for intervention effect over 1000 simulation runs using full data, CRA and cluster mean imputation for missing values under MAR.

k	m	ρ	Full data		CRA			LMM with cluster mean imputation
			Cluster level $t-$test	LMM approach	Cluster level $t-$test	Adjusted $t-$test	LMM approach	
5	30	0.01	63.8	60.9	49.3	49.3	47.7	49.6
		0.05	38.0	38.2	32.9	32.8	33.2	32.9
		0.10	25.6	26.6	23.2	23.6	24.1	23.2
	50	0.01	76.6	74.8	66.3	66.3	66.9	66.3
		0.05	42.9	39.6	39.5	39.8	39.3	39.5
		0.10	28.3	25.0	27.6	27.8	27.7	27.6
	100	0.01	90.5	89.2	82.8	82.5	83.0	82.8
		0.05	48.7	46.8	43.8	43.7	43.9	43.8
		0.10	28.7	29.3	27.0	27.2	27.2	27.0
	250	0.01	97.3	97.2	95.3	95.4	95.2	95.3
		0.05	54.2	51.7	51.9	51.5	51.8	51.9
		0.10	31.6	33.0	31.4	31.4	31.4	31.4
10	30	0.01	92.0	93.4	81.9	83.3	83.2	81.9
		0.05	68.2	71.0	61.2	62.3	61.9	61.2
		0.10	50.7	51.2	46.5	45.6	45.8	46.5
	50	0.01	98.4	97.7	95.2	95.2	95.2	95.2
		0.05	76.4	77.6	70.6	70.5	71.0	70.6
		0.10	52.1	54.8	51.0	49.8	50.6	51.0
	100	0.01	99.7	99.8	99.7	99.7	99.8	99.7
		0.05	82.7	84.1	81.0	80.8	81.0	81.0
		0.10	57.3	58.9	55.2	54.4	54.8	55.2
20	30	0.01	99.8	100	98.1	98.3	98.3	98.1
		0.05	94.3	96.0	88.8	89.0	89.4	88.8
		0.10	80.0	80.9	75.5	74.4	75.2	75.5
	50	0.01	100	100	99.9	99.9	99.9	99.9
		0.05	97.0	97.1	95.8	95.8	95.8	95.8
		0.10	84.0	86.1	83.3	83.4	83.5	83.0
	100	0.01	100	100	100	100	100	100
		0.05	98.4	98.9	97.9	97.8	97.8	97.9
		0.10	87.3	87.0	85.5	85.7	85.7	85.5

Appendix B: R code for estimating SEs and 95 % CIs of KM estimates using Williams approach

```
library(survival)
SE.Williams<-function(dataSet, ncls){
## dataSet: Data with the followings variables
        # obstime: minimum of survival time and censoring time
        # status: 1 for event and 0 for censored
        # clud.id: cluster id
## ncls: number of cluster

# estimating survival funciton
survf<-survfit(Surv(obstime, status)~ 1, conf.type="none", data=dataSet)
summ<-as.data.frame(summary(survf)[c(2,3,4,6,8)])
q.prob<-summ$n.event/summ$n.risk
var.williams<-array(NA,length(summ$time))

for(j in 1:length(summ$time)){
dataSet$delta<-0
dataSet$delta[dataSet$obstime==summ$time[j]]<-1
dataSet$gamma<-0
dataSet$gamma[dataSet$obstime>=summ$time[j]]<-1

# linearised value of q_{ij} for (ij)th subject (i-clustrer, j-individual)
dataSet$linear1<-(dataSet$delta - q.prob[j]*dataSet$gamma)/summ$n.risk[j]
# linearised value of S(t_v) for (ij)th subject (i-clustrer, j-individual)
```

```
if (j==1) {
dataSet$linear2<-dataSet$linear1
csum.linear2<-as.vector(with(dataSet,sapply(split(linear2,clus.id),sum)))
var.williams[j]<-ncls*var(csum.linear2)
} else {
dataSet$linear2<-(1-q.prob[j])*dataSet$linear2
                          +summ$surv[(j-1)]*dataSet$linear1
csum.linear2<-as.vector(with(dataSet,sapply(split(linear2,clus.id),sum)))
var.williams[j]<-ncls*var(csum.linear2)
}  }
# SEs of survival probability using Wiliams approach
summ$std.err.W<-sqrt(var.williams)
names(summ)[c(4,5)]<-c("surv.prob","std.err.G")



###95% CI for S(t) using complementary log-log transformation  ######
 if(conf.int==TRUE){
 den<-(summ$surv^2)*((log(summ$surv))^2)
 se.clogW<-sqrt(var.williams/den)

  ##confidence limits using Willaims SE
 summ$loweCI<- (summ$surv)^(exp(1.96*se.clogW))
 summ$upperCI<- (summ$surv)^(exp(-1.96*se.clogW))
 }
return(summ)
# time: event time points
# n.risk: number of individuals in the risk set
```

```
# n.event: number of events

# surv.prob: Estimated survival probabilities

# std.err.G: SEs of survival probabilities using Greenwood formula

# std.err.W: SEs of survival probabilities using Williams approach

}
```